(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **3 025 782**

(51) Int. Cl.:

G16B 25/20 (2009.01)
G16B 30/00 (2009.01)
G16B 30/10 (2009.01)
C12Q 1/6869 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **05.01.2018 E 21172159 (2)**

(97) Fecha y número de publicación de la concesión europea: **05.03.2025 EP 3889962**

(54) Título: **Métodos y sistemas para la generación de conjuntos de índices moleculares únicos con longitudes moleculares heterogéneas y la corrección de errores**

(30) Prioridad:

18.01.2017 US 201762447851 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
09.06.2025

(73) Titular/es:

ILLUMINA, INC. (100.00%)
5200 Illumina Way
San Diego, CA 92122, US

(72) Inventor/es:

WU, KEVIN;
ZHAO, CHEN;
CHUANG, HAN-YU;
SO, ALEX;
TANNER, STEPHEN y
GROSS, STEPHEN, M

(74) Agente/Representante:

DEL VALLE VALIENTE, Sonia

DESCRIPCIÓN

Métodos y sistemas para la generación de conjuntos de índices moleculares únicos con longitudes moleculares heterogéneas y la corrección de errores

Antecedentes

La tecnología de secuenciación de nueva generación proporciona una velocidad de secuenciación cada vez mayor, lo que permite una mayor profundidad de secuenciación. Sin embargo, dado que la precisión y la sensibilidad de la secuenciación se ven afectadas por errores y ruido de diversas fuentes, p. ej., defectos de la muestra, PCR durante la preparación de la genoteca, enriquecimiento, agrupación y secuenciación, el aumento de la profundidad de la secuenciación por sí solo no puede garantizar la detección de secuencias de muy baja frecuencia alélica, tales como en el ADN libre circulante (ADNlc) fetal en el plasma materno, el ADN tumoral circulante (ADNtc) y las mutaciones subclonales en agentes patógenos. Por lo tanto, es deseable desarrollar métodos para determinar secuencias de moléculas de ADN en pequeña cantidad y/o baja frecuencia alélica, suprimiendo al mismo tiempo la imprecisión de la secuenciación debida a diversas fuentes de error.

La patente US 2016/319345 A1 (GNERRE SANTE [US] Y COL.) expone métodos, sistemas y programas informáticos para determinar secuencias de fragmentos de ácido nucleico usando índices moleculares únicos (UMI). Los métodos de secuenciación determinan las secuencias de fragmentos de ácido nucleico de ambas cadenas de los fragmentos de ácido nucleico. Los métodos emplean UMI físicos ubicados en una o ambas cadenas de adaptadores de secuenciación. No se exponen el uso de UMI no aleatorios de longitud variable (vNRUMI) ni la determinación de puntuaciones de alineación que indiquen la similitud entre una subsecuencia de una lectura y un vNRUMI, en donde las puntuaciones de alineación se basan en emparejamientos de nucleótidos y modificaciones de nucleótidos entre la subsecuencia de la lectura y el vNRUMI.

Resumen

La presente invención se define mediante reivindicaciones independientes adjuntas. Las realizaciones preferidas se definen en las reivindicaciones dependientes.

Las implementaciones expuestas atañen a métodos, aparatos, sistemas y productos de programas informáticos para determinar secuencias de fragmentos de ácido nucleico utilizando índices moleculares únicos (UMI). En algunas implementaciones, los UMI incluyen UMI no aleatorios (NRUMI) o índices moleculares únicos no aleatorios y de longitud variable (vNRUMI).

Un aspecto de la exposición proporciona métodos para secuenciar moléculas de ácido nucleico a partir de una muestra. El método incluye: (a) aplicar adaptadores a los fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador incluye un índice molecular único no aleatorio y en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos no aleatorios y de longitud variable (vNRUMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI; (d) identificar, entre la pluralidad de lecturas, las lecturas asociadas con un mismo índice molecular único no aleatorio de longitud variable (vNRUMI); y (e) determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI.

En algunas implementaciones, la identificación de las lecturas asociadas con el mismo vNRUMI incluye obtener, para cada lectura de la pluralidad de lecturas, puntuaciones de alineación con respecto al conjunto de vNRUMI, indicando cada puntuación de alineación la similitud entre una subsecuencia de una lectura y un vNRUMI, en donde la subsecuencia está en una región de la lectura en la que probablemente se ubiquen los nucleótidos derivados del vNRUMI.

En algunas implementaciones, las puntuaciones de alineación se basan en los emparejamientos de nucleótidos y en las modificaciones de nucleótidos entre la subsecuencia de la lectura y el vNRUMI. En algunas implementaciones, las modificaciones de nucleótidos incluyen sustituciones, adiciones y deleciones de nucleótidos. En algunas implementaciones, cada puntuación de alineación penaliza los emparejamientos erróneos al principio de una secuencia, pero no penaliza los emparejamientos erróneos al final de la secuencia.

En algunas implementaciones, obtener una puntuación de alineación entre una lectura y un vNRUMI incluye: (a) calcular una puntuación de alineación entre el vNRUMI y cada una de las posibles secuencias de prefijo de la subsecuencia de la lectura; (b) calcular una puntuación de alineación entre la subsecuencia de la lectura y cada una de las posibles secuencias de prefijo del vNRUMI; y (c) obtener una puntuación de alineación más alta entre las puntuaciones de alineación calculadas en (a) y (b) como la puntuación de alineación entre la lectura y el vNRUMI.

En algunas implementaciones, la subsecuencia tiene una longitud que es igual a la longitud del vNRUMI más largo del conjunto de vNRUMI. En algunas implementaciones, identificar las lecturas asociadas con el mismo vNRUMI de (d) incluye además: seleccionar, para cada lectura de la pluralidad de lecturas, al menos un vNRUMI del conjunto de vNRUMI basándose en las puntuaciones de alineación; y asociar cada lectura de la pluralidad de lecturas con al menos el vNRUMI seleccionado para la lectura.

En algunas implementaciones, seleccionar al menos un vNRUMI del conjunto de vNRUMI incluye seleccionar un vNRUMI que tenga una puntuación de alineación más alta de entre el conjunto de vNRUMI. En algunas implementaciones, al menos un vNRUMI incluye dos o más vNRUMI.

En algunas implementaciones, el método incluye además seleccionar uno de los dos o más vNRUMI como el mismo vNRUMI de (d) y (e).

En algunas implementaciones, los adaptadores aplicados en (a) se obtienen mediante: (i) proporcionar un conjunto de secuencias de oligonucleótidos que tienen al menos dos longitudes moleculares diferentes; (ii) seleccionar un subconjunto de secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos, todas las distancias de modificación entre las secuencias de oligonucleótidos del subconjunto de secuencias de oligonucleótidos que cumplen un valor umbral, el subconjunto de secuencias de oligonucleótidos que forman el conjunto de vNRUMI; y (iii) sintetizar los adaptadores, cada uno de los cuales incluye una región hibridada bicatenaria, un brazo 5' monocatenario, un brazo 3' monocatenario y al menos un vNRUMI del conjunto de vNRUMI. En algunas implementaciones, el valor umbral es 3. En algunas implementaciones, el conjunto de vNRUMI incluye vNRUMI de 6 nucleótidos y vNRUMI de 7 nucleótidos.

En algunas implementaciones, la determinación de (e) incluye agrupar las lecturas asociadas con el mismo vNRUMI en un grupo para obtener una secuencia de nucleótidos de consenso para la secuencia del fragmento de ADN de la muestra. En algunas implementaciones, la secuencia de nucleótidos de consenso se obtiene basándose parcialmente en las puntuaciones de calidad de las lecturas.

En algunas implementaciones, la determinación de (e) incluye: identificar, entre las lecturas asociadas con el mismo vNRUMI, las lecturas que tienen una misma posición de lectura o posiciones de lectura similares en una secuencia de referencia, y determinar la secuencia del fragmento de ADN usando lecturas que (i) están asociadas con el mismo vNRUMI y (ii) tienen la misma posición de lectura o posiciones de lectura similares en la secuencia de referencia.

En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 10 000 vNRUMI diferentes. En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 1000 vNRUMI diferentes. En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 200 vNRUMI diferentes.

En algunas implementaciones, la aplicación de adaptadores a los fragmentos de ADN de la muestra incluye la aplicación de adaptadores a ambos extremos de los fragmentos de ADN de la muestra.

Otro aspecto de la exposición se refiere a los métodos para preparar adaptadores de secuenciación, donde los métodos incluyen: (a) proporcionar un conjunto de secuencias de oligonucleótidos que tienen al menos dos longitudes moleculares diferentes; (b) seleccionar un subconjunto de secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos, todas las distancias de modificación entre las secuencias de oligonucleótidos del subconjunto de secuencias de oligonucleótidos que cumplen un valor umbral, el subconjunto de secuencias de oligonucleótidos que forman un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI); y (c) sintetizar una pluralidad de adaptadores de secuenciación, en donde cada adaptador de secuenciación incluye una región hibridada bicatenaria, un brazo 5' monocatenario, un brazo 3' monocatenario y al menos un vNRUMI del conjunto de vNRUMI.

En algunas implementaciones, (b) incluye: (i) seleccionar una secuencia de oligonucleótidos del conjunto de secuencias de oligonucleótidos; (ii) añadir el oligonucleótido seleccionado a un conjunto creciente de secuencias de oligonucleótidos y eliminar el oligonucleótido seleccionado del conjunto de secuencias de oligonucleótidos para obtener un conjunto reducido de secuencias de oligonucleótidos; (iii) seleccionar una secuencia de oligonucleótidos instantánea del conjunto reducido que maximice una función de distancia, en donde la función de distancia es una distancia de modificación mínima entre la secuencia de oligonucleótidos instantánea y cualquier secuencia de oligonucleótidos del conjunto en expansión, y en donde la función de distancia alcance el valor umbral; (iv) añadir el oligonucleótido instantáneo al conjunto en expansión y eliminar el oligonucleótido instantáneo del conjunto reducido; (v) repetir (iii) y (iv) una o más veces; y (vi) proporcionar el conjunto en expansión como el subconjunto de secuencias de oligonucleótidos que forman el conjunto de vNRUMI.

En algunas implementaciones, (v) incluye repetir (iii) y (iv) hasta que la función de distancia ya no alcance el valor umbral.

En algunas implementaciones, (v) incluye repetir (iii) y (iv) hasta que el conjunto en expansión alcance un tamaño definido.

En algunas implementaciones, la secuencia de oligonucleótidos instantánea o una secuencia de oligonucleótidos del conjunto en expansión es más corta que una secuencia de oligonucleótidos más larga del conjunto de secuencias de oligonucleótidos. El método incluye además, antes de (iii), (1) incorporar una base de timina o una base de timina más cualquiera de las cuatro bases a la secuencia de oligonucleótidos instantánea o la secuencia de oligonucleótidos del conjunto en expansión, generando de esta manera una secuencia rellena que tiene la misma longitud que la secuencia de oligonucleótidos más larga del conjunto de secuencias de oligonucleótidos, y (2) usar la secuencia rellena para calcular la distancia mínima de modificación. En algunas implementaciones, las distancias de edición son distancias de Levenshtein. En algunas implementaciones, el valor umbral es 3.

En algunas implementaciones, el método incluye además, antes de (b), eliminar determinadas secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos para obtener un conjunto filtrado de secuencias de oligonucleótidos; y proporcionar el conjunto filtrado de secuencias de oligonucleótidos como el conjunto de secuencias de oligonucleótidos del que se selecciona el subconjunto.

En algunas implementaciones, las determinadas secuencias de oligonucleótidos incluyen secuencias de oligonucleótidos que tienen tres o más bases idénticas consecutivas. En algunas implementaciones, las determinadas secuencias de oligonucleótidos incluyen secuencias de oligonucleótidos que tienen un número combinado de bases de guanina y citosina menor que 2 y secuencias de oligonucleótidos que tienen un número combinado de bases de guanina y citosina mayor que 4.

En algunas implementaciones, las determinadas secuencias de oligonucleótidos incluyen secuencias de oligonucleótidos que tienen una misma base en las dos últimas posiciones. En algunas implementaciones, las determinadas secuencias de oligonucleótidos incluyen secuencias de oligonucleótidos que tienen una subsecuencia que coincide con el extremo 3' de uno o más cebadores de secuenciación.

En algunas implementaciones, las determinadas secuencias de oligonucleótidos incluyen secuencias de oligonucleótidos que tienen una base de timina en la última posición de las secuencias de oligonucleótidos.

En algunas implementaciones, el conjunto de vNRUMI incluye vNRUMI de 6 nucleótidos y vNRUMI de 7 nucleótidos.

Un aspecto adicional de la exposición se refiere a un método para secuenciar moléculas de ácido nucleico de una muestra, incluido (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador incluye un índice molecular único no aleatorio, y en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI; y (d) identificar, entre la pluralidad de lecturas, las lecturas asociadas con un mismo índice molecular único no aleatorio de longitud variable (vNRUMI).

En algunas implementaciones, el método además incluye obtener un recuento de las lecturas asociadas con el mismo vNRUMI.

Otro aspecto de la exposición se refiere a un método para secuenciar moléculas de ácido nucleico de una muestra, incluido (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador incluye un índice molecular único (UMI) y en donde los índices moleculares únicos (UMI) de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos de longitud variable (vUMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vUMI; y (d) identificar, entre la pluralidad de lecturas, las lecturas asociadas con un mismo índice molecular único de longitud variable (vUMI).

En algunas implementaciones, el método además incluye determinar una secuencia de un fragmento de ADN de la muestra usando las lecturas asociadas con el mismo vUMI.

En algunas implementaciones, el método además incluye obtener un recuento de las lecturas asociadas con los mismos vUMI.

Otro aspecto más de la exposición se refiere al método para secuenciar moléculas de ácido nucleico de una muestra, incluido (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador incluye un índice molecular único (UMI) de un conjunto de índices moleculares únicos (UMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de UMI; (d) obtener, para cada lectura de la pluralidad de lecturas, puntuaciones de alineación con respecto al conjunto de UMI, indicando cada puntuación de alineación la similitud entre una subsecuencia de una

lectura y un UMI; (e) identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo UMI utilizando las puntuaciones de alineación; y (e) determinar una secuencia de un fragmento de ADN en la muestra utilizando las lecturas asociadas con el mismo UMI.

En algunas implementaciones, las puntuaciones de alineación se basan en los emparejamientos de nucleótidos y en las modificaciones de nucleótidos entre la subsecuencia de la lectura y el UMI. En algunas implementaciones, cada puntuación de alineación penaliza los emparejamientos erróneos al principio de una secuencia, pero no penaliza los emparejamientos erróneos al final de la secuencia. En algunas implementaciones, el conjunto de UMI incluye UMI de al menos dos longitudes moleculares diferentes.

También se exponen sistemas, aparatos y productos de programas informáticos para determinar secuencias de fragmentos de ADN implementando los métodos expuestos.

Un aspecto de la exposición proporciona un producto de programa informático que incluye un medio legible por ordenador no transitorio que almacena código de programa que, cuando lo ejecuta uno o más procesadores de un sistema informático, hace que el sistema informático implemente un método para determinar la información de secuencia de una secuencia de interés de una muestra utilizando índices moleculares únicos (UMI). El código del programa incluye instrucciones para realizar los métodos anteriores.

Aunque los ejemplos de la presente memoria se refieren a seres humanos y el lenguaje está dirigido principalmente a problemas humanos, los conceptos descritos en la presente memoria pueden aplicarse a ácidos nucleicos de cualquier virus, planta, animal u otro organismo, y a poblaciones de los mismos (metagenomas, poblaciones víricas, etc.). Estas y otras características de la presente exposición se harán totalmente evidentes a partir de la siguiente descripción, con referencia a las figuras, y las reivindicaciones adjuntas, o pueden aprenderse mediante la puesta en práctica de la exposición como se indica a continuación en la memoria.

Breve descripción de las figuras

La Figura 1A es un diagrama de flujo que ilustra un ejemplo de flujo de trabajo utilizando UMI para secuenciar fragmentos de ácido nucleico.

La Figura 1B muestra un fragmento/molécula de ADN y los adaptadores empleados en las etapas iniciales del flujo de trabajo mostrado en la Figura 1A.

La Figura 1C es un diagrama de bloques que muestra un proceso para secuenciar fragmentos de ADN utilizando vNRUMI para suprimir errores.

La Figura 1D ilustra un proceso 140 para crear adaptadores de secuenciación que tengan vNRUMI.

La Figura 1E muestra ejemplos de cómo se puede comparar una subsecuencia de una secuencia de lectura o consulta (Q) con dos secuencias de referencia (S1 y S2) en el conjunto de vNRUMI.

La Figura 1F ilustra ejemplos de cómo las puntuaciones de alineación glocales pueden proporcionar una mejor supresión de errores que las puntuaciones de alineación globales.

La Figura 2A ilustra esquemáticamente cinco diseños de adaptador diferentes que pueden adoptarse en las diversas implementaciones.

La Figura 2B ilustra un proceso hipotético en el que se produce el salto de UMI en una reacción de PCR que implica adaptadores que tienen dos UMI físicos en dos brazos.

La Figura 2C muestra datos que contrastan las puntuaciones de calidad de lectura de las lecturas de secuencia utilizando NRUMI contra una condición de control.

Las Figuras 3A y 3B son diagramas que muestran los materiales y productos de reacción de los adaptadores de unión a fragmentos bicatenarios de acuerdo con algunos métodos expuestos en la presente memoria.

Las Figuras 4A-4E ilustran cómo los métodos divulgados en la presente memoria pueden suprimir diferentes fuentes de error en la determinación de la secuencia de un fragmento de ADN bicatenario. La Figura 5 ilustra esquemáticamente la aplicación de UMI físicos y UMI virtuales para obtener eficientemente lecturas largas de ambos extremos.

La Figura 6 es un diagrama de bloques de un sistema disperso para procesar una muestra de prueba.

La Figura 7 ilustra un sistema informático que puede servir como aparato computacional de acuerdo con determinadas realizaciones.

Descripción detallada

La exposición se refiere a métodos, aparatos, sistemas y productos de programas informáticos para secuenciar ácidos nucleicos, especialmente ácidos nucleicos con cantidad limitada o baja concentración, tal como ADNlc fetal en plasma materno o ADN tumoral circulante (ADNtc) en la sangre de un paciente con cáncer.

Los intervalos numéricos incluyen los números que definen el intervalo. Se pretende que cada limitación numérica máxima dada a lo largo de esta memoria descriptiva incluya cada limitación numérica inferior, como si tales limitaciones numéricas inferiores se escribieran expresamente en la presente memoria. Cada limitación numérica mínima dada a lo largo de esta memoria descriptiva incluirá cada limitación numérica más alta, como si tales limitaciones numéricas superiores estuvieran expresamente escritas en la presente memoria. Cada intervalo numérico dado a lo largo de esta memoria descriptiva incluirá cada intervalo numérico más estrecho que se encuentre dentro de tal intervalo numérico más amplio, como si tales intervalos numéricos más estrechos estuvieran expresamente escritos en la presente memoria.

Salvo que se defina de cualquier otra manera en la presente memoria, todos los términos técnicos y científicos utilizados en la presente memoria tienen el significado que entiende comúnmente un experto en la técnica. Diversos diccionarios científicos que incluyen los términos incluidos en la presente memoria son bien conocidos y están disponibles para los expertos en la técnica. Aunque cualquier método y materiales similares o equivalentes a los descritos en la presente memoria encuentran uso en la práctica o prueba de las realizaciones expuestas en la presente memoria, se describen algunos métodos y materiales.

Los términos definidos inmediatamente a continuación se describen más totalmente por referencia a la memoria descriptiva en su conjunto. Debe entenderse que la invención no se limita a la metodología, los protocolos y los reactivos particulares descritos, ya que estos pueden variar, dependiendo del contexto en el que los utilicen los expertos en la técnica.

Definiciones

Como se utilizan en la presente memoria, los términos en singular “un”, “una” y “el” o “la” incluyen la referencia al plural, a menos que el contexto lo indique de cualquier otra manera.

Salvo que se indique lo contrario, los ácidos nucleicos se escriben de izquierda a derecha en orientación de 5' a 3' y las secuencias de aminoácidos se escriben de izquierda a derecha en orientación de amino a carboxi, respectivamente.

Los índices moleculares únicos (UMI) son secuencias de nucleótidos aplicadas o identificadas en moléculas de ADN que pueden utilizarse para distinguir moléculas de ADN individuales entre sí. Dado que los UMI se utilizan para identificar moléculas de ADN, también se denominan identificadores moleculares únicos. Véase, p. ej., Kivioja, Nature Methods 9, 72–74 (2012). Los UMI pueden secuenciarse junto con las moléculas de ADN con las que están asociados para determinar si las secuencias leídas son las de una molécula de ADN fuente u otra. El término “UMI” se utiliza en la presente memoria para referirse tanto a la información de secuencia de un polinucleótido como al polinucleótido físico *per se.*

Normalmente, se secuencian múltiples instancias de una misma molécula fuente. En el caso de la secuenciación por síntesis mediante la tecnología de secuenciación de Illumina, la molécula fuente puede amplificarse mediante la PCR antes de aplicarlas a una cubeta de lectura. Amplificadas o no por la PCR, las moléculas individuales de ADN aplicadas a la cubeta de lectura se amplifican por puente o ExAmp para producir un grupo. Cada molécula de un grupo procede de la misma molécula de ADN de origen, pero se secuencian por separado. Para la corrección de errores y otros fines, puede ser importante determinar que todas las lecturas de un mismo grupo se identifiquen como derivadas de la misma molécula de origen. Los UMI permiten esta agrupación. Una molécula de ADN que se copia por amplificación o de cualquier otra manera para producir múltiples instancias de la molécula de ADN se denomina molécula de ADN fuente.

Además de los errores asociados con las moléculas de ADN fuente, también pueden producirse errores en una región asociada a los UMI. En algunas implementaciones, este último tipo de error puede corregirse mapeando una secuencia de lectura con un UMI más probable entre un grupo de UMI.

Los UMI son similares a los códigos de barras, que se suelen utilizar para distinguir las lecturas de una muestra de las lecturas de otras muestras, pero los UMI se utilizan para distinguir una molécula de ADN fuente de otra cuando se secuencian muchas moléculas de ADN juntas. Dado que puede haber muchas más moléculas de ADN en una muestra que muestras en un experimento de secuenciación, suele haber muchos más UMI distintos que códigos de barras distintos en un experimento de secuenciación.

Como ya se ha mencionado, los UMI pueden aplicarse o identificarse en moléculas de ADN individuales. En algunas implementaciones, los UMI pueden aplicarse a las moléculas de ADN mediante métodos que enlazan o unen

físicamente los UMI a las moléculas de ADN, p. ej., mediante unión o transposición a través de polimerasas, endonucleasas, transposasas, etc. Por lo tanto, estos UMI "aplicados" también se denominan UMI físicos. En algunos contextos, también pueden denominarse UMI exógenos. Los UMI identificados dentro de las moléculas de ADN fuente se denominan UMI virtuales. En algunos contextos, los UMI virtuales también pueden denominarse UMI endógenos.

Los UMI físicos pueden definirse de muchas maneras. Por ejemplo, pueden ser secuencias de nucleótidos aleatorias, pseudoaleatorias o parcialmente aleatorias, o no aleatorias, que se introducen en adaptadores o se incorporan de cualquier otra manera a las moléculas de ADN fuente que se van a secuenciar. En algunas implementaciones, los UMI físicos pueden ser tan únicos que se espera que cada uno de ellos identifique de forma única cualquier molécula de ADN fuente concreta presente en una muestra. Se genera la colección de adaptadores, cada una con un UMI físico, y esos adaptadores se unen a fragmentos u otras moléculas de ADN fuente que se van a secuenciar, y cada una de las moléculas secuenciadas individuales tiene un UMI que ayuda a distinguirlas de todos los demás fragmentos. En tales implementaciones, puede utilizarse un gran número de diferentes UMI físicos (p. ej., de varios miles a millones) para identificar de manera única fragmentos de ADN en una muestra.

Por supuesto, el UMI físico debe tener una longitud suficiente para garantizar esta unicidad para todas y cada una de las moléculas de ADN fuente. En algunas implementaciones, se puede utilizar un identificador molecular menos único junto con otras técnicas de identificación para garantizar que cada molécula de ADN fuente se identifique de forma única durante el proceso de secuenciación. En tales implementaciones, múltiples fragmentos o adaptadores pueden tener el mismo UMI físico. Otra información, tal como la ubicación de la alineación o los UMI virtuales, puede combinarse con el UMI físico para identificar de forma exclusiva las lecturas como derivadas de una única molécula/fragmento de ADN fuente. En algunas implementaciones, los adaptadores incluyen UMI físicos limitados a un número relativamente pequeño de secuencias no aleatorias, p. ej., 120 secuencias no aleatorias. Estos UMI físicos también se denominan UMI no aleatorios. En algunas implementaciones, los UMI no aleatorios pueden combinarse con información de posición de secuencia, la posición de secuencia y/o UMI virtuales para identificar lecturas atribuibles a una misma molécula de ADN fuente. Las lecturas identificadas pueden combinarse para obtener una secuencia de consenso que refleje la secuencia de la molécula de ADN fuente, como se describe en la presente memoria. Mediante el uso de UMI físicos, UMI virtuales y/o ubicaciones de alineación, se pueden identificar las lecturas que tienen UMI o ubicaciones iguales o relacionados, y las lecturas identificadas se pueden combinar después para obtener una o más secuencias de consenso. El proceso de combinar lecturas para obtener una secuencia de consenso también se denomina lecturas "de agrupación", que se describe además a continuación en la memoria.

Un "índice molecular único virtual" o "UMI virtual" es una subsecuencia única de una molécula de ADN fuente. En algunas implementaciones, los UMI virtuales se localizan en o cerca de los extremos de la molécula de ADN fuente. Una o más de estas posiciones finales únicas pueden, por sí solas o junto con otra información, identificar de forma única una molécula de ADN fuente. Dependiendo del número de moléculas de ADN fuente distintas y del número de nucleótidos en el UMI virtual, uno o más UMI virtuales pueden identificar de forma única moléculas de ADN fuente en una muestra. En algunos casos, se requiere una combinación de dos identificadores moleculares únicos virtuales para identificar una molécula de ADN fuente. Tales combinaciones pueden ser extremadamente raras, posiblemente encontradas solo una vez en una muestra. En algunos casos, uno o más UMI virtuales en combinación con uno o más UMI físicos pueden identificar conjuntamente de forma única una molécula de ADN fuente.

Un "UMI aleatorio" puede considerarse un UMI físico seleccionado como muestra aleatoria, con o sin reemplazo, de un conjunto de UMI que consiste en todas las posibles secuencias de oligonucleótidos diferentes dadas una o más longitudes de secuencia. Por ejemplo, si cada UMI del conjunto de UMI tiene n nucleótidos, el conjunto incluye 4^n UMI que tienen secuencias diferentes entre sí. Una muestra aleatoria seleccionada entre los 4^n UMI constituye un UMI aleatorio.

Por el contrario, un "UMI no aleatorio" (NRUMI), como se utiliza en la presente memoria, se refiere a un UMI físico que no es un UMI aleatorio. En algunas realizaciones, los UMI no aleatorios están predefinidos para un experimento o aplicación concretos. En ciertas realizaciones, se utilizan reglas para generar secuencias para un conjunto o seleccionar una muestra del conjunto para obtener un UMI no aleatorio. Por ejemplo, las secuencias de un conjunto pueden generarse de tal manera que las secuencias tengan un patrón o patrones particulares. En algunas implementaciones, cada secuencia difiere de cualquier otra secuencia del conjunto en un número determinado de nucleótidos (p. ej., 2, 3 o 4). Es decir, ninguna secuencia de UMI no aleatorio puede convertirse en otra secuencia de UMI no aleatorio disponible sustituyendo menos que el número determinado de nucleótidos. En algunas implementaciones, un conjunto de NRUMI utilizados en un proceso de secuenciación incluye menos que todos los posibles UMI dada una longitud de secuencia particular. Por ejemplo, un conjunto de NRUMI que tenga 6 nucleótidos puede incluir un total de 96 secuencias diferentes, en lugar de un total de 4^6=4096 posibles secuencias diferentes.

En algunas implementaciones donde los UMI no aleatorios se seleccionan entre un conjunto con menos de todas las secuencias diferentes posibles, el número de UMI no aleatorios es menor, a veces significativamente menor, que el número de moléculas de ADN fuente. En tales implementaciones, la información de UMI no aleatorios puede combinarse con otra información, tales como UMI virtuales, ubicaciones de lectura en una secuencia de referencia y/o información de secuencia de lecturas, para identificar lecturas de secuencias derivadas de una misma molécula de ADN fuente.

El término "índice molecular no aleatorio de longitud variable" (vNRUMI) se refiere a un UMI de un conjunto de vNRUMI seleccionado de entre un grupo de UMI de longitudes moleculares variables (o longitud heterogénea) mediante un proceso de selección no aleatorio. El término vNRUMI se usa para referirse tanto a la molécula del UMI así como a la secuencia del UMI. En algunas implementaciones, determinados UMI pueden eliminarse del grupo de UMI para proporcionar un grupo filtrado de UMI, grupo que después se usa para generar el conjunto de vNRUMI.

En algunas implementaciones, cada vNRUMI difiere de todos los demás vNRUMI del conjunto utilizado en un proceso en al menos una distancia de modificación definida. En algunas implementaciones, un conjunto de vNRUMI usado en un proceso de secuenciación incluye menos que todos los UMI posibles dadas las longitudes moleculares relevantes. Por ejemplo, un conjunto de vNRUMI que tenga 6 y 7 nucleótidos puede incluir un total de 120 secuencias diferentes (en lugar de un total de 4^6+4^7=20 480 posibles secuencias diferentes). En otras implementaciones, las secuencias no se seleccionan aleatoriamente de un conjunto. En cambio, algunas secuencias se seleccionan con mayor probabilidad que otras.

El término "longitud molecular" también se denomina longitud de secuencia y se puede medir en nucleótidos. El término "longitud molecular" también se usa indistintamente con los términos tamaño molecular, tamaño del ADN y longitud de la secuencia.

La distancia de edición es una métrica que cuantifica la diferencia entre dos cadenas (por ejemplo, palabras) contando el número mínimo de operaciones necesarias para transformar una cadena en la otra. En bioinformática, se puede usar para cuantificar la similitud de las secuencias de ADN, que se pueden ver como cadenas de las letras A, C, G y T.

Las diferentes formas de distancia de edición usan diferentes conjuntos de operaciones de cadena. La distancia de Levenshtein es un tipo común de distancia de edición. Las operaciones de cadena de la distancia de Levenshtein tienen en cuenta el número de eliminaciones, inserciones y sustituciones de caracteres en la cadena. En algunas implementaciones, se pueden usar otras variantes de distancias de edición. Por ejemplo, se pueden obtener otras variantes de la distancia de edición restringiendo el conjunto de operaciones. La distancia de subsecuencia común más larga (LCS) es la distancia de edición, con la inserción y la eliminación como las dos únicas operaciones de edición, ambas con un coste por unidad. Del mismo modo, al permitir solo sustituciones, se obtiene la distancia de Hamming, que está restringida a cadenas de igual longitud. La distancia de Jaro-Winkler se puede obtener a partir de una distancia de edición donde solo se permiten las transposiciones.

En algunas implementaciones, las diferentes operaciones de cadena se pueden ponderar de manera diferente para una distancia de edición. Por ejemplo, una operación de sustitución puede ponderarse con un valor de 3, mientras que un indel puede ponderarse con un valor de 2. En algunas implementaciones, las coincidencias de diferentes tipos pueden ponderarse de diferente modo. Por ejemplo, una coincidencia A-A podría tener el doble de ponderación que una coincidencia G-G.

Una puntuación de alineación es una puntuación que indica una similitud de dos secuencias determinada usando un método de alineación. En algunas implementaciones, una puntuación de alineación tiene en cuenta el número de modificaciones (p. ej., deleciones, inserciones y sustituciones de caracteres en la cadena). En algunas implementaciones, una puntuación de alineación tiene en cuenta un número de emparejamientos. En algunas implementaciones, una puntuación de alineación tiene en cuenta tanto el número de emparejamientos como el número de modificaciones. En algunas implementaciones, el número de emparejamientos y modificaciones se pondera por igual para la puntuación de alineación. Por ejemplo, una puntuación de alineación se puede calcular de la siguiente manera: *n.º de emparejamientos - n.º de inserciones - n.º de deleciones - n.º de sustituciones*. En otras implementaciones, los números de emparejamientos y modificaciones se pueden ponderar de manera diferente. Por ejemplo, una puntuación de alineación se puede calcular de la siguiente manera: *n.º de emparejamientos* × *5* - n.º *de inserciones* × *4* - *n.º de deleciones* × *4* - *n.º de sustituciones* × *6.*

La expresión "lecturas de ambos extremos" se refiere a lecturas obtenidas mediante secuenciación de ambos extremos, que obtiene una lectura de cada uno de los extremos de un fragmento de ácido nucleico. La secuenciación de ambos extremos implica fragmentar el ADN en secuencias denominadas insertos. En algunos protocolos tales como algunos usados por Illumina, las lecturas de insertos más cortos (p. ej., del orden de decenas a cientos de pb) se denominan lecturas de ambos extremos de inserto corto o simplemente lecturas de ambos extremos. Por el contrario, las lecturas de insertos más largos (p. ej., del orden de varios miles de pb) se denominan lecturas de pares de parejas. En la presente exposición, pueden utilizarse lecturas de ambos extremos de inserto corto y lecturas de pares de parejas de inserto largo y estas no se diferencian con respecto al proceso para determinar secuencias de fragmentos de ADN. Por lo tanto, la expresión "lectura de ambos extremos" puede referirse tanto a lecturas de ambos extremos de inserto corto como a lecturas de pares de parejas de inserto largo, que se describen en más detalle después en la presente memoria. En algunas realizaciones, las lecturas de ambos extremos incluyen lecturas de aproximadamente 20 pb a 1000 pb. En algunas realizaciones, las lecturas de ambos extremos incluyen lecturas de aproximadamente 50 pb a 500 pb, de aproximadamente 80 pb a 150 pb, o de aproximadamente 100 pb.

Como se utiliza en la presente memoria, los términos "alineación" y "alinear" se refieren al proceso de comparar una lectura con una secuencia de referencia y, de esta manera, determinar si la secuencia de referencia contiene la secuencia leída. Un proceso de alineamiento, como se utiliza en la presente memoria, intenta determinar si una lectura se puede mapear a una secuencia de referencia, pero no siempre resulta en una lectura alineada con la secuencia de referencia. Si la secuencia de referencia contiene la lectura, la lectura puede cartografiarse a la secuencia de referencia o, en determinadas realizaciones, a una ubicación particular en la secuencia de referencia. En algunos casos, el alineamiento simplemente dice si una lectura es o no un miembro de una secuencia de referencia particular (es decir, si la lectura está presente o ausente en la secuencia de referencia). Por ejemplo, el alineamiento de una lectura con la secuencia de referencia para el cromosoma 13 humano indicará si la lectura está presente en la secuencia de referencia para el cromosoma 13.

Por supuesto, las herramientas de alineación tienen muchos aspectos adicionales y muchas otras aplicaciones en bioinformática que no se describen en esta solicitud. Por ejemplo, las alineaciones también se pueden usar para determinar cuán similares son dos secuencias de ADN de dos especies diferentes, proporcionando así una medida de cuán estrechamente relacionadas están en un árbol evolutivo.

En algunas implementaciones de la presente memoria, la alineación se realiza entre una subsecuencia de una lectura y un vNRUMI como secuencia de referencia para determinar una puntuación de alineación como se describe además después en la presente memoria. Las puntuaciones de alineación entre una lectura y múltiples vNRUMI se pueden usar después para determinar a cuál de los vNRUMI se debe asociar o mapear la lectura.

En algunos casos, un alineamiento indica adicionalmente una ubicación en la secuencia de referencia donde la lectura se cartografía. Por ejemplo, si la secuencia de referencia es la secuencia del genoma humano completo, un alineamiento puede indicar que una lectura está presente en el cromosoma 13, y puede indicar además que la lectura está en una cadena y/o sitio particular del cromosoma 13. En algunas situaciones, las herramientas de alineamiento son imperfectas, en el sentido de que a) no se encuentran todas las alineaciones válidas y b) algunas alineaciones obtenidas son inválidas. Esto ocurre por diversas razones, p. ej., las lecturas pueden contener errores y las lecturas secuenciadas pueden ser diferentes del genoma de referencia debido a diferencias de haplotipo. En algunas aplicaciones, las herramientas de alineación incluyen tolerancia a emparejamientos erróneos integrada, lo que tolera ciertos grados de emparejamientos erróneos de pares de bases y aun así permite la alineación de las lecturas con una secuencia de referencia. Esto puede ayudar a identificar alineaciones válidas de lecturas que de cualquier otra manera pasarían desapercibidas.

Las lecturas alineadas son una o más secuencias que se identifican como una coincidencia en términos del orden de sus moléculas de ácido nucleico a una secuencia de referencia conocida, tal como un genoma de referencia. Una lectura alineada y su ubicación determinada en la secuencia de referencia constituyen una marcador de secuencia. El alineamiento puede realizarse manualmente, aunque típicamente se implementa mediante un algoritmo informático, ya que sería imposible alinear las lecturas en un período de tiempo razonable para implementar los métodos expuestos en la presente memoria. Un ejemplo de un algoritmo para alinear secuencias es el método de alineación híbrido global-local (glocal) para comparar una secuencia de prefijo de una lectura con un vNRUMI como se describe además a continuación en la memoria. Otro ejemplo de un método de alineación es el programa informático Efficient Local Alignment of Nucleotide Data (ELAND) distribuido como parte del flujo de trabajo de análisis genómico de Illumina. Como alternativa, se puede emplear un filtro de Bloom o un analizador de pertenencia de conjunto similar para alinear las lecturas con los genomas de referencia. Véase la solicitud de patente de Estados Unidos n.º 14/354,528, presentada el 25 de abril de 2014, que se menciona en la presente memoria como referencia en su totalidad. La coincidencia de una lectura de secuencia en alineación puede ser una coincidencia de secuencia del 100 % o menos del 100 % (es decir, una coincidencia no perfecta). Se exponen métodos de alineación adicionales en la solicitud de patente US-15/130,668 (referencia legal ILMNP008) presentada el 15 de abril de 2016, que se menciona como referencia en su totalidad.

El término "cartografía" usado en la presente memoria se refiere a asignar una secuencia leída a una secuencia más grande, p. ej., un genoma de referencia, mediante alineación.

Los términos y expresiones "polinucleótido", "ácido nucleico" y "moléculas de ácido nucleico" se usan indistintamente y hacen referencia a una secuencia de nucleótidos unidos covalentemente (es decir, ribonucleótidos en el caso del ARN y desoxirribonucleótidos en el caso del ADN), en los que la posición 3' de la pentosa de un nucleótido está unida mediante un grupo fosfodiéster a la posición 5' de la pentosa del siguiente. Los nucleótidos incluyen secuencias de cualquier forma de ácido nucleico, incluidas, aunque no de forma limitativa, moléculas de ARN y ADN tales como moléculas de ADN libre circulante (ADNlc). El término "polinucleótido" incluye, sin limitarse a, polinucleótidos monocatenarios y bicatenarios.

La expresión "muestra de ensayo" en la presente memoria se refiere a una muestra, típicamente obtenida de un líquido biológico, célula, tejido, órgano u organismo, que incluye un ácido nucleico o una mezcla de ácidos nucleicos que tienen al menos una secuencia de ácido nucleico que se va a analizar en cuanto a la variación del número de copias y otras alteraciones genéticas, tales como, aunque no de forma limitativa, polimorfismo de un solo nucleótido, inserciones, deleciones y variaciones estructurales. En determinadas realizaciones, la muestra tiene al menos una

secuencia de ácido nucleico cuyo número de copias se sospecha que tiene una variación. Tales muestras incluyen, aunque no de forma limitativa, esputo/fluido oral, líquido amniótico, sangre, una fracción de sangre, o muestras de biopsia con aguja fina, orina, líquido peritoneal, líquido pleural y similares. Aunque la muestra frecuentemente se toma de un sujeto humano (p. ej., un paciente), los ensayos pueden utilizarse con muestras de cualquier mamífero, lo que incluye, aunque no de forma limitativa, perros, gatos, caballos, cabras, ovejas, ganado bovino, cerdos, etc., así como poblaciones mixtas, tales como poblaciones microbianas silvestres o poblaciones de virus de pacientes. La muestra puede utilizarse directamente tal como se obtiene de la fuente biológica o tras un pretratamiento para modificar el carácter de la muestra. Por ejemplo, tal pretratamiento puede incluir preparar plasma a partir de sangre, diluir fluidos viscosos, etc. Los métodos de pretratamiento también pueden implicar, aunque no de forma limitativa, filtración, precipitación, dilución, destilación, mezcla, centrifugación, congelación, liofilización, concentración, amplificación, fragmentación de ácido nucleico, inactivación de componentes interferentes, la adición de reactivos, lisis, etc. Si tales métodos de pretratamiento se emplean con respecto a la muestra, tales métodos de pretratamiento son típicamente tales que el(los) ácido(s) nucleico(s) de interés permanecen en la muestra de prueba, a veces en una concentración proporcional a la de una muestra de prueba no tratada (p. ej., especialmente, una muestra que no se somete a ninguno de tal(es) método(s) de pretratamiento). Tales muestras "tratadas" o "procesadas" siguen siendo consideradas muestras biológicas "de ensayo" con respecto a los métodos descritos en la presente memoria.

La expresión "secuenciación de nueva generación (NGS)" en la presente memoria se refiere a métodos de secuenciación que permiten la secuenciación masiva en paralelo de moléculas amplificadas clonalmente y de moléculas de ácido nucleico individuales. Los ejemplos no limitativos de NGS incluyen secuenciación por síntesis usando terminadores de colorante reversibles y secuenciación por ligadura.

El término "lectura" se refiere a una lectura de secuencia de una porción de una muestra de ácido nucleico. De forma típica, aunque no necesariamente, una lectura representa una secuencia corta de pares de bases contiguos en la muestra. La lectura puede representarse simbólicamente mediante la secuencia de pares de bases en A, T, C y G de la porción de muestra, junto con una estimación probabilística de la corrección de la base (puntuación de calidad). Puede almacenarse en un dispositivo de memoria y procesarse según sea apropiado para determinar si coincide con una secuencia de referencia o cumple otros criterios. Se puede obtener una lectura directamente de un aparato de secuenciación o indirectamente a partir de información de secuencia almacenada con respecto a la muestra. En algunos casos, una lectura es una secuencia de ADN de longitud suficiente (p. ej., al menos aproximadamente 20 pb) que puede usarse para identificar una secuencia o región más grande, p. ej., que puede alinearse y mapearse a un cromosoma o región genómica o gen.

Los términos "sitio" y "ubicación de la alineación" se utilizan indistintamente para hacer referencia a una posición única (es decir, ID del cromosoma, posición en el cromosoma y orientación) en un genoma de referencia. En algunas realizaciones, un sitio puede ser un residuo, un marcador de secuencia o una posición de segmento en una secuencia de referencia.

Como se utiliza en la presente memoria, la expresión "genoma de referencia" o "secuencia de referencia" se refiere a cualquier secuencia genética conocida concreta, ya sea parcial o completa, de cualquier organismo o virus, que pueda usarse para referenciar secuencias identificadas de un sujeto. Por ejemplo, un genoma de referencia utilizado para sujetos humanos, así como muchos organismos diferentes, se encuentra en el National Center for Biotechnology Information en ncbi.nlm.nih.gov. Un "genoma" se refiere a la información genética completa de un organismo o virus, expresada en secuencias de ácido nucleico. Sin embargo, se entiende que "completo" es un concepto relativo, ya que aun en el genoma de referencia estándar se espera incluir huecos y errores.

En algunas implementaciones, se puede usar una secuencia de vNRUMI como secuencia de referencia con la que se alinea una secuencia de prefijo de una lectura. La alineación proporciona una puntuación de alineación entre la secuencia de prefijo de la lectura y el vNRUMI, que se puede utilizar para determinar si la lectura y el vNRUMI deben asociarse en un proceso de agrupar las lecturas asociadas con el mismo vNRUMI.

En diversas realizaciones, la secuencia de referencia es significativamente más grande que las lecturas con las que están alineadas. Por ejemplo, puede ser al menos aproximadamente 100 veces mayor, o al menos aproximadamente 1000 veces mayor, o al menos aproximadamente 10.000 veces mayor, o al menos aproximadamente 10^5 veces mayor, o al menos aproximadamente 10^6 veces mayor, o al menos aproximadamente 10^7 veces mayor.

En un ejemplo, la secuencia de referencia es la de un genoma humano de longitud completa. Tales secuencias pueden denominarse secuencias de referencia genómicas. En otro ejemplo, la secuencia de referencia se limita a un cromosoma humano específico tal como el cromosoma 13. En algunas realizaciones, un cromosoma Y de referencia es la secuencia cromosómica Y de la versión del genoma humano hg19. Tales secuencias pueden denominarse secuencias de referencia cromosómicas. Otros ejemplos de secuencias de referencia incluyen genomas de otras especies, así como cromosomas, regiones subcromosómicas (tales como cadenas), etc., de cualquier especie.

En algunas realizaciones, una secuencia de referencia para el alineamiento puede tener una longitud de secuencia de aproximadamente 1 a aproximadamente 100 veces la longitud de una lectura. En tales realizaciones, la alineación y la secuenciación se consideran una alineación o secuenciación dirigidas, en lugar de una alineación o secuenciación

del genoma completo. En estas realizaciones, la secuencia de referencia normalmente incluye una secuencia génica y/u otra secuencia restringida de interés. En este sentido, la alineación de una subsecuencia de una lectura con un vNRUMI es una forma de alineación dirigida.

En diversas realizaciones, la secuencia de referencia es una secuencia de consenso u otra combinación derivada de múltiples individuos. Sin embargo, en determinadas aplicaciones, la secuencia de referencia puede tomarse de un determinado individuo.

La expresión "obtenido de" cuando se usa en el contexto de un ácido nucleico o una mezcla de ácidos nucleicos, en la presente memoria se refiere a los medios donde se obtienen el(los) ácido(s) nucleico(s) de la fuente en la que se originan. Por ejemplo, en una realización, una mezcla de ácidos nucleicos que se deriva de dos genomas diferentes significa que los ácidos nucleicos, p. ej., ADNlc, fueron liberados naturalmente por las células a través de procesos naturales tales como necrosis o apoptosis. En otra realización, una mezcla de ácidos nucleicos que se deriva de dos genomas diferentes significa que los ácidos nucleicos se extrajeron de dos tipos diferentes de células de un sujeto.

La expresión "líquido biológico" en la presente memoria se refiere a un líquido tomado de una fuente biológica e incluye, por ejemplo, sangre, suero, plasma, esputo, líquido de lavado, líquido cefalorraquídeo, orina, semen, sudor, lágrimas, saliva y similares. Como se utilizan en la presente memoria, los términos "sangre", "plasma" y "suero" abarcan expresamente fracciones o porciones procesadas de los mismos. Similarmente, cuando una muestra se toma de una biopsia, un hisopo, un frotis, etc., la "muestra" abarca expresamente una fracción o porción procesada obtenida de la biopsia, el hisopo, el frotis, etc.

Como se utiliza en la presente memoria, el término "cromosoma" se refiere al portador hereditario del gene que porta una célula viva, que se deriva de cadenas de cromatina que comprenden componentes de ADN y proteínas (especialmente histonas). En la presente memoria se utiliza el sistema convencional de numeración de cromosomas del genoma humano individual reconocido internacionalmente.

Como se utiliza en la presente memoria, la expresión "cebador" se refiere a un oligonucleótido aislado que es capaz de actuar como punto de inicio de la síntesis cuando se somete a condiciones que inducen la síntesis de un producto de extensión (p. ej., las condiciones incluyen nucleótidos, un agente inductor, tal como ADN-polimerasa, los iones y moléculas necesarios y una temperatura y pH adecuados). El cebador puede ser preferiblemente monocatenario para la máxima eficiencia en la amplificación, pero como alternativa puede ser bicatenario. Si es bicatenario, el cebador se trata primero para separar sus cadenas antes de usarse para preparar productos de extensión. El cebador puede ser un oligodesoxirribonucleótido. El cebador es suficientemente largo para cebar la síntesis de productos de extensión en presencia del agente inductor. Las longitudes exactas de los cebadores dependerán de muchos factores, incluida la temperatura, la fuente del cebador, el uso del método y los parámetros utilizados para el diseño del cebador.

Introducción y contexto

La tecnología de secuenciación de nueva generación (NGS) se ha desarrollado rápidamente, proporcionando nuevas herramientas para avanzar en la investigación y la ciencia, así como en la asistencia sanitaria y los servicios que dependen de la información genética y biológica relacionada. Los métodos de NGS se realizan de forma masiva en paralelo, lo que ofrece una velocidad cada vez mayor para determinar la información de secuencia de biomoléculas. Sin embargo, muchos de los métodos de NGS y las técnicas de manipulación de muestras asociadas introducen errores, de modo que las secuencias resultantes tienen una tasa de error relativamente alta, que oscila entre un error en unos pocos cientos de pares de bases hasta un error en unos pocos miles de pares de bases. Estas tasas de error son a veces aceptables para determinar la información genética heredable, tal como las mutaciones de la línea germinal, porque tal información es coherente en la mayoría de las células somáticas, que proporcionan muchas copias del mismo genoma en una muestra de ensayo. Un error originado por la lectura de una copia de una secuencia tiene un impacto menor o eliminable cuando se leen sin error muchas copias de la misma secuencia. Por ejemplo, si una lectura errónea de una copia de una secuencia no puede alinearse correctamente con una secuencia de referencia, puede simplemente desecharse del análisis. Las lecturas sin errores de otras copias de la misma secuencia pueden seguir proporcionando información suficiente para realizar análisis válidos. Como alternativa, en lugar de desechar la lectura que tenga un par de bases diferente a otras lecturas de la misma secuencia, se puede descartar el par de bases diferente como resultado de una fuente de error conocida o desconocida.

Sin embargo, tales planteamientos de corrección no tienen un buen funcionamiento para detectar secuencias con bajas frecuencias alélicas, tales como mutaciones somáticas subclonales encontradas en los ácidos nucleicos de tejido tumoral, ADN tumoral circulante, ADNlc fetal de baja concentración en el plasma materno, mutaciones farmacorresistentes de agentes patógenos, etc. En estos ejemplos, un fragmento de ADN puede portar una mutación somática de interés en un sitio de secuencia, mientras que muchos otros fragmentos en el mismo sitio de secuencia no tienen la mutación de interés. En tal escenario, las lecturas de secuencia o pares de bases del fragmento de ADN mutado podrían no utilizarse o malinterpretarse en la secuenciación convencional, de esta manera se perdería información para detectar la mutación de interés.

Debido a estas diversas fuentes de error, el aumento de la profundidad de secuenciación por sí solo no puede garantizar la detección de variaciones somáticas con una frecuencia alélica muy baja (p. ej., <1 %). Algunas implementaciones divulgadas en la presente memoria proporcionan métodos de secuenciación dúplex que suprimen eficazmente los errores en situaciones en las que las señales de secuencias válidas de interés son bajas, tales como muestras con bajas frecuencias alélicas.

Los índices moleculares únicos (UMI) permiten el uso de información de múltiples lecturas para suprimir el ruido de secuenciación. Los UMI, junto con la información contextual, tal como las posiciones de alineación, nos permiten rastrear el origen de cada lectura hasta una molécula de ADN original específica. Dadas las múltiples lecturas producidas por la misma molécula de ADN, se pueden utilizar planteamientos computacionales para separar las variantes reales (es decir, las variantes biológicamente presentes en las moléculas de ADN originales) de las variantes introducidas artificialmente mediante un error de secuenciación. Las variantes pueden incluir, aunque no de forma limitativa, inserciones, deleciones, variantes de múltiples nucleótidos, variantes de un solo nucleótido y variantes estructurales. Con esta información, podemos inferir la verdadera secuencia de las moléculas de ADN. Nos referimos a esta metodología computacional como agrupación de lecturas. Esta tecnología de reducción de errores tiene varias aplicaciones importantes. En el contexto del análisis de ADN libre circulante, las variantes importantes frecuentemente se dan con frecuencias extremadamente bajas (es decir, <1 %); por lo tanto, su señal puede quedar anulada por errores de secuenciación. La reducción de ruido basada en el UMI nos permite llamar con mucha más precisión a estas variantes de baja frecuencia. Los UMI y la agrupación de lecturas también pueden ayudar a identificar los duplicados de PCR en datos de alta cobertura, lo que permite mediciones de frecuencia de variantes más precisas.

En algunas implementaciones, se usan UMI aleatorios, en los que se unió una secuencia aleatoria a las moléculas de ADN, y esas secuencias aleatorias se usaron como códigos de barras de UMI. Sin embargo, el uso de un conjunto de UMI no aleatorios diseñados a propósito permitió una fabricación más sencilla en algunas implementaciones. Dado que este planteamiento no es aleatorio, los UMI reciben el nombre de UMI no aleatorios (NRUMI). En algunas implementaciones, un conjunto de NRUMI consiste en secuencias de longitud uniforme (p. ej., $n = 6$ nucleótidos de longitud). Debido al proceso A-tailing mediante el cual estas moléculas de NRUMI se ligan a las moléculas de ADN, la 7.ª lectura ($n + 1$) es invariablemente una timina (T). Esta uniformidad puede provocar una degradación en la calidad de lectura que se propaga a lo largo de los ciclos de lectura aguas abajo de esta base. Este efecto se ilustra en la Figura 2C.

Aunque este problema puede ser menos prominente en las cubetas de lectura sin patrón secuenciadas con 4 colorantes, es probable que su gravedad aumente en las cubetas de lectura con patrón secuenciadas con 2 colorantes, ya que la llamada de bases se vuelve inherentemente más difícil. En algunas implementaciones, se usa un proceso novedoso para generar conjuntos de NRUMI de longitudes mixtas, identificando de manera única tales NRUMI de longitud variable (vNRUMI) y corrigiendo los errores dentro de estos vNRUMI. Ofrece diversidad en la generación y distinción de códigos de barras de ADN de longitud heterogénea. Los resultados experimentales muestran que el método vNRUMI es más robusto (es decir, más capaz de corregir los errores de secuenciación) que las soluciones convencionales.

En algunas implementaciones, se usa un algoritmo ambicioso para construir de forma iterativa conjuntos de vNRUMI. En cada iteración, selecciona una secuencia de un grupo de candidatos de vNRUMI tal que la secuencia elegida maximice la distancia mínima de Levenshtein entre ella y cualquier vNRUMI que ya se haya elegido. Si múltiples secuencias comparten el valor máximo de esta métrica, el algoritmo elige una de esas secuencias al azar, prefiriendo las secuencias de menor longitud. Esta métrica de distancia debe ser de al menos 3 para garantizar una buena corrección de errores dentro del conjunto de vNRUMI resultante; si no se puede satisfacer esta condición, el proceso detiene la adición de nuevos vNRUMI al conjunto y devuelve el conjunto tal cual. Todo este proceso se puede repetir para generar diferentes conjuntos de vNRUMI con características similares.

Los adaptadores pueden incluir UMI físicos que permitan a uno determinar de qué cadena del fragmento de ADN proceden las lecturas. Algunas realizaciones aprovechan esto para determinar una primera secuencia consenso para las lecturas procedentes de una cadena del fragmento de ADN, y una segunda secuencia consenso para la cadena complementaria. En muchas realizaciones, una secuencia de consenso incluye los nucleótidos detectados en todas o en la mayor parte de las lecturas, excluyendo a su vez los nucleótidos que aparecen en pocas de las lecturas. Pueden aplicarse distintos criterios de consenso. El proceso de combinar lecturas basadas en UMI o ubicaciones de alineamiento para obtener una secuencia de consenso también se denomina "agrupación" de lecturas. Utilizando UMI físicos, UMI virtuales y/o ubicaciones de alineamiento, se puede determinar que las lecturas para las secuencias consenso primera y segunda proceden del mismo fragmento bicatenario. Por lo tanto, en algunas realizaciones, se determina una tercera secuencia de consenso utilizando las secuencias de consenso primera y segunda obtenidas para la misma molécula/fragmento de ADN, incluida la tercera secuencia de consenso nucleótidos comunes para las secuencias de consenso primera y segunda, mientras se excluyen aquellas incoherentes entre las dos. En implementaciones alternativas, solo se obtiene directamente una secuencia de consenso agrupando todas las lecturas derivadas de ambas cadenas del mismo fragmento, en lugar de comparando las dos secuencias de consenso obtenidas a partir de las dos cadenas. Por último, la secuencia del fragmento puede determinarse a partir de la tercera o la única secuencia consenso, que incluye pares de bases que son coincidentes en las lecturas procedentes de ambas cadenas del fragmento.

En algunas realizaciones, el método combina diferentes tipos de índices para determinar el polinucleótido fuente del que se derivan las lecturas. Por ejemplo, el método puede utilizar UMI físicos y virtuales para identificar lecturas derivadas de una única molécula de ADN. Al utilizar una segunda forma de UMI, además del UMI físico, los UMI físicos pueden ser más cortos que cuando solo se utilizan UMI físicos para determinar el polinucleótido fuente. Esta estrategia tiene un impacto mínimo en el rendimiento de la preparación de bibliotecas y no requiere una longitud de lectura de secuenciación adicional.

Algunas aplicaciones de los métodos divulgados incluyen:

- Supresión de errores para la detección de mutaciones somáticas. Por ejemplo, la detección de mutaciones con una frecuencia alélica de menos del 0,1 % es muy importante en la biopsia líquida del ADN tumoral circulante.

- Corregir los errores de prefase, fase y otros errores de secuenciación para lograr lecturas largas de alta calidad (p. ej., 1x1000 pb)

- Reducir el tiempo de ciclo para una longitud de lectura fija y corregir el incremento de la fase y la prefase con este método.

- Usar UMI a ambos lados del fragmento para crear lecturas de ambos extremos largas virtuales. Por ejemplo, unir una lectura de 2×500 realizando 500+50 en duplicados.

- Cuantificar o contar los fragmentos de ácido nucleico relacionados con una secuencia de interés.

Flujo de trabajo para la secuenciación de fragmentos de ácido nucleico mediante UMI

La Figura 1A es un diagrama de flujo que ilustra un ejemplo de flujo de trabajo 100 para utilizar UMI para secuenciar fragmentos de ácido nucleico. El flujo de trabajo 100 es ilustrativos solo de algunas implementaciones. Se entiende que algunas implementaciones emplean flujos de trabajo con operaciones adicionales no ilustradas aquí, mientras que otras implementaciones pueden omitir algunas de las operaciones ilustradas aquí. Por ejemplo, algunas implementaciones no requieren la operación 102 y/o la operación 104. También, el flujo de trabajo 100 se emplea para la secuenciación del genoma completo. En algunas implementaciones que implican la secuenciación dirigida, se pueden aplicar etapas operativas para hibridar y enriquecer ciertas regiones entre las operaciones 110 y 112.

La operación 102 proporciona fragmentos de ADN bicatenario. Los fragmentos de ADN pueden obtenerse, por ejemplo, fragmentando ADN genómico, recogiendo ADN fragmentado de forma natural (p. ej., ADNlc o ADNtc) o sintetizando fragmentos de ADN a partir de ARN. En algunas implementaciones, para sintetizar fragmentos de ADN a partir de ARN, ARN mensajero o ARN no codificante, primero se purifica el ARN mensajero usando selección con poliA o agotamiento del ARN ribosómico y, a continuación, el ARNm seleccionado se fragmenta químicamente y se convierte en ADNc monocatenario usando cebado hexámero aleatorio. Se genera una cadena complementaria del ADNc para crear un ADNc bicatenario que está listo para la construcción de bibliotecas. Para obtener fragmentos de ADN bicatenario a partir de ADN genómico (ADNg), el ADNg de entrada se fragmenta, p. ej., mediante cizallamiento hidrodinámico, nebulización, fragmentación enzimática, etc., para generar fragmentos de longitudes adecuadas, p. ej., de aproximadamente 1000 pb, 800 pb, 500 o 200 pb. Por ejemplo, la nebulización puede romper el ADN en trozos de menos de 800 pb en cortos periodos de tiempo. Este proceso genera fragmentos de ADN bicatenario.

En algunas implementaciones, el ADN fragmentado o dañado puede procesarse sin requerir una fragmentación adicional. Por ejemplo, el ADN fijado con formalina y embebido en parafina (FFPE) o cierto cfADN a veces están lo suficientemente fragmentados como para que no se requiera una etapa de fragmentación adicional.

La Figura 1B muestra un fragmento/molécula de ADN y los adaptadores empleados en las etapas iniciales del flujo de trabajo 100 en la Figura 1A. Aunque en la Figura 1B solo se ilustra un fragmento bicatenario, en el flujo de trabajo se pueden preparar simultáneamente de miles a millones de fragmentos de una muestra. La fragmentación del ADN por métodos físicos produce extremos heterogéneos, que comprenden una mezcla de salientes 3', salientes 5' y extremos romos. Los salientes serán de longitudes variables y los extremos pueden estar fosforilados o no. Un ejemplo de los fragmentos de ADN bicatenario obtenidos a partir de la fragmentación del ADN genómico de la operación 102 se muestra como fragmento 123 en la Figura 1B.

El fragmento 123 tiene tanto un saliente 3' en el extremo izquierdo como un saliente 5' en el extremo derecho, y está marcado con ρ y ϕ, lo que indica dos secuencias en el fragmento que pueden utilizarse como UMI virtuales en algunas implementaciones, que, cuando se utilizan solas o combinadas con UMI físicos de un adaptador que se ligará al fragmento, pueden identificar de forma única el fragmento. Los UMI se asocian de forma exclusiva a un único fragmento de ADN de una muestra que incluye un polinucleótido fuente y su cadena complementaria. Un UMI físico es una secuencia de un oligonucleótido unido al polinucleótido fuente, a su cadena complementaria o a un polinucleótido procedente del polinucleótido fuente. Un UMI virtual es una secuencia de un oligonucleótido dentro del

polinucleótido fuente, su cadena complementaria o un polinucleótido procedente del polinucleótido fuente. Dentro de este esquema, también se puede hacer referencia al UMI físico como un UMI extrínseco o exógeno, y al UMI virtual como un UMI intrínseco o endógeno.

Las dos secuencias ρ y ϕ se refieren en realidad a dos secuencias complementarias en el mismo sitio genómico, pero para simplificar, se indican en una sola cadena en algunos de los fragmentos bicatenarios que se muestran en la presente memoria. Los UMI virtuales como ρ y ϕ pueden utilizarse en un paso posterior del flujo de trabajo para ayudar a identificar las lecturas que se originan en una o ambas cadenas del fragmento fuente de ADN único. Una vez identificadas las lecturas, pueden agruparse para obtener una secuencia consenso.

Si los fragmentos de ADN se producen por métodos físicos, el flujo de trabajo 100 procede a realizar la operación de reparación del extremo 104, que produce fragmentos de extremos romos con extremos fosforilados en 5'. En algunas implementaciones, este paso convierte los salientes resultantes de la fragmentación en extremos romos utilizando ADN polimerasa T4 y enzima Klenow. La actividad exonucleasa de 3' a 5' de estas enzimas elimina los salientes 3' y la actividad polimerasa de 5' a 3' rellena los salientes 5'. Además, la polinucleótido quinasa T4 fosforila en esta reacción los extremos 5' de los fragmentos de ADN. El fragmento 125 de la Figura 1B es un ejemplo de producto con extremos reparados y romos.

Después de la reparación de los extremos, el flujo de trabajo 100 procede a la operación 106 para adenilar los extremos 3' de los fragmentos, lo que también se denomina A-tailing o dA-tailing, ya que se añade un único dATP a los extremos 3' de los fragmentos romos para evitar que se liguen entre sí durante la reacción de ligadura de adaptadores. La molécula bicatenaria 127 de la Figura 1B muestra un fragmento con adición de cola de A con extremos romos con salientes 3'-dA y extremos 5'-fosfato. Un solo nucleótido "T" en el extremo 3' de cada uno de los dos adaptadores de secuenciación, como se ve en el punto 129 de la Figura 1B, proporciona un saliente complementario al saliente 3'-dA en cada extremo del inserto para ligar los dos adaptadores al inserto.

Después de adenilar los extremos 3', el flujo de trabajo 100 procede a la operación 108 para ligar adaptadores parcialmente bicatenarios a ambos extremos de los fragmentos. En algunas implementaciones, los adaptadores utilizados en una reacción incluyen diferentes UMI físicos para asociar las lecturas de secuencia a un único polinucleótido fuente, que puede ser un fragmento de ADN monocatenario o bicatenario. En algunas implementaciones, un conjunto de UMI físicos usados en una reacción son UMI aleatorios. En algunas implementaciones, el conjunto de UMI físicos usados en una reacción son UMI no aleatorios (NRUMI). En algunas implementaciones, el conjunto de UMI físicos usados en una reacción son UMI no aleatorios de longitud variable (vNRUMI).

El elemento 129 de la Figura 1B ilustra dos adaptadores que deben ligarse al fragmento bicatenario que incluye dos UMI virtuales ρ y ϕ cerca de los extremos del fragmento. Estos adaptadores se ilustran basándose en los adaptadores de secuenciación de la plataforma de Illumina, ya que diversas implementaciones pueden utilizar la plataforma de NGS de Illumina para obtener lecturas y detectar secuencias de interés. El adaptador mostrado a la izquierda incluye el UMI físico α en su región bicatenaria, mientras que el adaptador de la derecha incluye el UMI físico β en su región bicatenaria. En la cadena que tiene el extremo 5' desnaturalizado, en dirección de 5' a 3', los adaptadores tienen una secuencia P5, una secuencia de índice y una secuencia con 2 cebadores de lectura (α o β). En la cadena que tiene el extremo 3' desnaturalizado, en dirección de 3' a 5', los adaptadores tienen una secuencia P7', una secuencia de índice, una secuencia con 1 cebador de lectura y el UMI físico (α o β).

Los oligonucleótidos P5 y P7' son complementarios a los cebadores de amplificación unidos a la superficie de las cubetas de lectura de la plataforma de secuenciación de Illumina. En algunas implementaciones, la secuencia de índice proporciona un medio para realizar un seguimiento de la fuente de una muestra, permitiendo así la multiplexación de múltiples muestras en la plataforma de secuenciación. Pueden utilizarse otros diseños de adaptadores y plataformas de secuenciación en diversas implementaciones. Los adaptadores y la tecnología de secuenciación se describen con más detalle en las secciones siguientes.

La reacción representada en la Figura 1B añade secuencias distintas al fragmento genómico. En la Figura 1B se ilustra un producto de ligamiento 120 del mismo fragmento descrito anteriormente. Este producto de ligadura 120 tiene el UMI físico α, el UMI virtual ρ, el UMI virtual ϕ y el UMI físico β en su cadena superior, en la dirección 5'-3'. El producto de ligadura también tiene el UMI físico β, el UMI virtual ϕ, el UMI virtual ρ y el UMI físico α en su cadena inferior, en la dirección 5'-3'. La presente exposición incorpora métodos que utilizan tecnologías de secuenciación y adaptadores distintos de los proporcionados por Illumina.

Aunque los adaptadores de ejemplo aquí tienen los UMI físicos en las regiones bicatenarias de los adaptadores, algunas implementaciones usan adaptadores que tienen UMI físicos en las regiones monocatenarias, tales como los adaptadores (i) y (iv) en las Figuras 2A.

En algunas implementaciones, los productos de esta reacción de ligamiento se purifican y/o se seleccionan por tamaño mediante electroforesis en gel de agarosa o bolas magnéticas. A continuación, el ADN seleccionado por tamaño se amplifica mediante la PCR para enriquecer los fragmentos que tienen adaptadores en ambos extremos. Véase el

bloque 110. Como se mencionó antes, en algunas implementaciones, se pueden aplicar las operaciones para hibridar y enriquecer ciertas regiones de los fragmentos de ADN para dirigirse a las regiones a secuenciar.

A continuación, el flujo de trabajo 100 procede a la amplificación en grupo de los productos de PCR, p. ej., en una plataforma de Illumina. Véase la operación 112. Mediante la agrupación de los productos de la PCR, las bibliotecas pueden agruparse para multiplexación, p. ej., con hasta 12 muestras por carril, utilizando diferentes secuencias de índices de adaptadores para realizar un seguimiento de las diferentes muestras.

Tras la amplificación del grupo, pueden obtenerse lecturas de secuenciación mediante secuenciación por síntesis en la plataforma de Illumina. Véase la operación 114. Aunque los adaptadores y el proceso de secuenciación descritos en la presente memoria se basan en la plataforma de Illumina, pueden utilizarse otras tecnologías de secuenciación, especialmente métodos de NGS, en lugar o además de la plataforma de Illumina.

El flujo de trabajo 100 puede agrupar las lecturas que tienen los mismos UMI físicos y/o los mismos UMI virtuales en uno o más grupos, obteniendo de esta manera una o más secuencias de consenso. Véase la operación 116. En algunas implementaciones, los UMI físicos son UMI aleatorios. En algunas implementaciones, los UMI físicos son UMI no aleatorios. En algunas implementaciones, los UMI físicos son UMI aleatorios de longitud variable. En algunas implementaciones, los UMI físicos son UMI no aleatorios de longitud variable (vNRUMI). Una secuencia consenso incluye bases de nucleótidos que son coherentes o cumplen un criterio de consenso entre lecturas en un grupo agrupado. En algunas implementaciones, los UMI físicos por sí solos pueden proporcionar información suficiente para etiquetar fragmentos de ADN y agrupar las lecturas. Tales implementaciones requerirían un número suficientemente grande de UMI físicos para etiquetar de forma única los fragmentos de ADN. En otras implementaciones, los UMI físicos, los UMI virtuales y la información de posición pueden combinarse de diversas formas para agrupar las lecturas y obtener secuencias de consenso para determinar la secuencia de un fragmento o al menos una porción del mismo. En algunas implementaciones, los UMI físicos se combinan con UMI virtuales para agrupar las lecturas. En otras implementaciones, los UMI físicos y las posiciones de lectura se combinan para agrupar las lecturas. La información sobre la posición de las lecturas puede obtenerse mediante diversas técnicas que utilizan diferentes medidas de posición, p. ej., coordenadas genómicas de las lecturas, posiciones en una secuencia de referencia o posiciones cromosómicas. En otras implementaciones, los UMI físicos, los UMI virtuales y las posiciones de lectura se combinan para agrupar las lecturas.

Finalmente, el flujo de trabajo 100 utiliza las una o más secuencias consenso para determinar la secuencia del fragmento de ácido nucleico de la muestra. Véase la operación 118. Esto puede implicar determinar la secuencia del fragmento de ácido nucleico como la tercera secuencia consenso o la secuencia consenso única descrita anteriormente.

En una implementación particular que incluye operaciones similares a las operaciones 108–119, un método para secuenciar moléculas de ácido nucleico de una muestra usando UMI no aleatorios implica lo siguiente: (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos de adaptadores de ADN, en donde cada adaptador comprende un NRUMI, y en donde los NRUMI de los adaptadores tienen al menos dos longitudes moleculares diferentes, formando un conjunto de vNRUMI; (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI; (b) identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo vNRUMI; y (e) determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI.

En otra implementación, se usan UMI aleatorios de longitud variable para secuenciar moléculas de ácido nucleico. El método incluye: (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador comprende un índice molecular único (UMI) y en donde los índices moleculares únicos (UMI) de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos de longitud variable (vUMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vUMI; y (d) identificar, entre la pluralidad de lecturas, las lecturas asociadas con un mismo índice molecular único no aleatorio de longitud variable (vNRUMI). Algunas implementaciones además incluyen determinar una secuencia de un fragmento de ADN de la muestra usando las lecturas asociadas con el mismo vUMI.

En algunas implementaciones, los UMI usados para secuenciar fragmentos de ácido nucleico pueden ser UMI aleatorios de longitud fija, UMI no aleatorios de longitud fija, UMI aleatorios de longitud variable, UMI no aleatorios de longitud variable o cualquier combinación de los mismos. En estas implementaciones, el método para secuenciar fragmentos de ácido nucleico incluye: (a) aplicar adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador comprende un índice molecular único (UMI) en un conjunto de índices moleculares únicos (UMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de UMI; (d) obtener, para cada lectura de la pluralidad de lecturas, puntuaciones de alineación con respecto al conjunto de UMI, indicando cada puntuación de alineación la similitud

entre una subsecuencia de una lectura y un UMI; (e) identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo UMI utilizando las puntuaciones de alineación; y (e) determinar una secuencia de un fragmento de ADN en la muestra utilizando las lecturas asociadas con el mismo UMI. En algunas implementaciones, las puntuaciones de alineación se basan en los emparejamientos de nucleótidos y en las modificaciones de nucleótidos entre la subsecuencia de la lectura y el UMI. En algunas implementaciones, cada puntuación de alineación penaliza los emparejamientos erróneos al principio de una secuencia, pero no penaliza los emparejamientos erróneos al final de la secuencia.

En algunas implementaciones, las lecturas de secuencia son lecturas de ambos extremos. Cada lectura o bien incluye un UMI no aleatorio o bien está asociada con un UMI no aleatorio a través de una lectura de ambos extremos. En algunas implementaciones, las longitudes de lectura son más cortas que los fragmentos de ADN o más cortas que la mitad de la longitud de los fragmentos. En estos casos, a veces no se determina la secuencia completa de todo el fragmento. En cambio, se determinan los dos extremos del fragmento. Por ejemplo, un fragmento de ADN puede ser de 500 pb de longitud, a partir del cual pueden obtenerse dos lecturas de ambos extremos de 100 pb. En este ejemplo, pueden determinarse las 100 bases de cada extremo del fragmento, y los 300 pb del centro del fragmento pueden no determinarse sin utilizar información de otras lecturas. En algunas implementaciones, si las dos lecturas en ambos extremos son lo suficientemente largas como para solaparse, puede determinarse la secuencia completa del fragmento completo a partir de las dos lecturas. Por ejemplo, véase el ejemplo descrito en relación con la Figura 5.

En algunas implementaciones, un adaptador tiene un UMI dúplex no aleatorio en la región de bicatenaria del adaptador, y cada lectura incluye un primer UMI no aleatorio en un extremo y un segundo UMI no aleatorio en el otro extremo.

Método para la secuenciación de fragmentos de ácido nucleico mediante vNRUMI

En algunas implementaciones, los vNRUMI se incorporan a adaptadores para la secuenciación de fragmentos de ADN. Los vNRUMI proporcionan un mecanismo para suprimir los diferentes tipos de errores que se producen en un flujo de trabajo tal como el descrito anteriormente. Algunos de los errores pueden producirse en la fase de procesamiento de la muestra, tales como deleciones, adiciones y sustituciones en el procesamiento de la muestra. Pueden producirse otros errores en la fase de secuenciación. Algunos errores pueden estar ubicados en las bases derivadas de los fragmentos de ADN, otros errores pueden estar ubicados en las bases correspondientes a los UMI de los adaptadores.

Algunas implementaciones proporcionan un proceso novedoso para detectar y corregir errores en los vNRUMI y en las lecturas de secuencia. En un nivel alto, dada una lectura que contiene un vNRUMI (potencialmente malinterpretado) y sus bases aguas abajo, el proceso utiliza una estrategia de alineación híbrida global-local (glocal) para emparejar las pocas primeras bases de la lectura con un vNRUMI conocido, obteniendo de esta manera puntuaciones de alineación entre las secuencias de prefijo de la lectura y el vNRUMI conocido. Se determina que un vNRUMI que tiene una puntuación de alineación glocal más alta es el vNRUMI asociado a la lectura, lo que proporciona un mecanismo para agrupar la lectura con otras lecturas asociadas con el mismo vNRUMI, corrigiendo de esta manera los errores. El pseudocódigo para obtener puntuaciones de alineación glocales y emparejar vNRUMI utilizando las puntuaciones de alineación glocales en algunas implementaciones se proporciona de la siguiente manera.

```
algoritmo glocal:
    entrada: Secuencias de ADN x e y
             Puntuaciones integrales para (emparejamiento, emparejamiento erróneo, hueco),
             predeterminadas (1, -1, -1)
    salida: z, un valor integral que aumenta con la similitud de secuencia
    puntuaciones = matriz numérica de longitud (x)+1 filas y longitud (y)+1 columnas

      para i de 0 a longitud(x), inclusive:
          puntuaciones [i][0] = i
      para j de 0 a longitud(y), inclusive:
          puntuaciones [0][j] = j
      para i de 1 a longitud(x), inclusive:
          para j de 1 a longitud(y), inclusive:
              coste = emparejamiento si x[i-1]==y[j-1], de cualquier otra manera, coste = emparejamiento
              erróneo
              establecer puntuaciones [i][j] hasta un máximo de:
                  puntuaciones [i-1][j-1] + coste
                  puntuaciones[i-1][j] + hueco
                  puntuaciones[i][j-1] + hueco
    z = máximo de la última fila y última columna de la matriz de puntuaciones
    devolver z

algoritmo match_vNRUMI:
```

entrada: conjunto X que contiene todos los vNRUMI válidos/no mutados
secuencia Q, un vNRUMI posiblemente mutado y bases aguas abajo
salida: m_1 el conjunto de coincidencias más probables de vNRUMI
m_2: el conjunto de las segundas coincidencias más parecidas a vNRUMI
potentialLengths = longitudes únicas de todas las secuencias de X
matchScores = lista que contiene posibles coincidencias para Q y sus puntuaciones correspondientes
n = longitud máxima de cualquier secuencia del conjunto X
subseq = primeras n bases en Q
para cada secuencia S en X:
registrar la puntuación glocal (S, subseq) en matchScores, junto con
la secuencia S en sí misma
m_1 = secuencias de X con las puntuaciones glocales más altas observadas
m_2 = secuencias de X con la segunda puntuación glocal más alta observada
devuelve m_1 y m_2

Vale la pena señalar el uso de una métrica de distancia no convencional. En otras metodologías comparables para códigos de barras de ADN, la mayoría adoptan heurísticas que cuantifican la distancia de modificación, especialmente, la distancia de Levenshtein, la distancia de Hamming o derivados de las mismas. Conceptualmente, una puntuación de alineación proporciona una métrica similar de similitud de secuencia, pero con una diferencia clave: cuenta los emparejamientos además de los cambios. Una heurística compatible con los emparejamientos detecta algunas de las ventajas de algunas implementaciones de los NRUMI de longitud variable.

En algunas implementaciones, no se usa ni una alineación global tradicional Needleman-Wunsch ni un método tradicional de alineación local Smith-Waterman, pero se usa un planteamiento híbrido novedoso. Especialmente, la alineación utiliza un planteamiento de Needleman-Wunsch al principio de la alineación, lo que penaliza las modificaciones allí realizadas, pero aprovecha los conceptos de la alineación local de Smith Waterman al final de la alineación al no penalizar las modificaciones finales. En este sentido, el planteamiento de alineación actual abarca tanto un componente global como uno local y, por lo tanto, se denomina planteamiento de alineación glocal. En el caso de un error de inserción o deleción en la secuenciación, la alineación cambiaría considerablemente. Este planteamiento glocal no penalizaría ese único evento más de lo que se penalizaría una mutación puntual única. Permitir los huecos intermedios nos permite lograrlo.

El planteamiento de alineación glocal tiene la capacidad de trabajar con grupos de códigos de barras de longitud heterogénea, una característica distintiva de las metodologías convencionales.

Al identificar emparejamientos, algunas implementaciones pueden devolver múltiples emparejamientos de vNRUMI como los "mejores" cuando hay empates. Aunque el pseudocódigo anterior solo refleja los mejores y los segundos mejores conjuntos devueltos, algunas implementaciones tienen la capacidad de devolver más de solo dos conjuntos de vNRUMI, tales como un segundo mejor conjunto, un tercer mejor conjunto, un cuarto mejor conjunto, etc. Al proporcionar más información de buenos emparejamientos, el proceso puede corregir mejor los errores al agrupar las lecturas asociadas con uno o más emparejamientos candidatos de vNRUMI. La Figura 1C es un diagrama de bloques que muestra un proceso para la secuenciación de fragmentos de ADN utilizando vNRUMI para suprimir los errores que se producen en los fragmentos de ADN y los errores de los UMI que se utilizan para etiquetar las moléculas fuente de los fragmentos de ADN. El proceso 130 comienza aplicando adaptadores a fragmentos de ADN en una muestra para obtener productos adaptadores de ADN. Véase el bloque 131. Cada adaptador de los adaptadores tiene un índice molecular único no aleatorio. Los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares no aleatorios de longitud variable (vNRUMI).

En algunas implementaciones, se une, liga, inserta, incorpora o vincula de cualquier otra manera un adaptador a cada extremo de los fragmentos de ADN. En algunas implementaciones, la muestra que contiene los fragmentos de ADN es una muestra de sangre. En algunas implementaciones, los fragmentos de ADN contienen fragmentos de ADN libres de células. En algunas implementaciones, los fragmentos de ADN incluyen ADN libre circulante que se origina en un tumor, y la secuencia de los fragmentos de ADN de la muestra es indicativa del tumor.

El proceso 130 procede amplificando los productos adaptadores de ADN para obtener una pluralidad de polinucleótidos amplificados. Véase el bloque 132. El proceso 130 implica además secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI. Véase el bloque 133. Además, el proceso 130 implica identificar lecturas asociadas con un mismo vNRUMI de entre la pluralidad de lecturas. Véase el bloque 134. Finalmente, el proceso 130 incluye determinar una secuencia de un fragmento de ADN de la muestra usando las lecturas asociadas con el mismo vNRUMI.

Como se mencionó anteriormente, el proceso 130 ilustrado en la Figura 1C proporciona un método para la secuenciación de fragmentos de ADN usando vNRUMI. El proceso 130 comienza aplicando adaptadores a fragmentos de ADN de la muestra para obtener productos adaptadores de ADN (bloque 131). El proceso 130 también implica la amplificación de los productos adaptadores de ADN para obtener una pluralidad de polinucleótidos amplificados

(bloque 132); la secuenciación de la calidad de los polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI (bloque 133); la identificación de las lecturas asociadas al mismo vNRUMI (bloque 134); y la determinación de una secuencia de fragmentos de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI (bloque 135). La muestra puede ser una muestra de sangre, una muestra de plasma, una muestra de tejido o una de las muestras como se describe en cualquier otro sitio de la presente memoria. En algunas implementaciones, los adaptadores aplicados en la etapa 131 se pueden obtener a partir de un proceso tal como el proceso 140 ilustrado en la Figura 1D.

En algunas implementaciones, los vNRUMI de los adaptadores tienen al menos dos longitudes moleculares diferentes. En algunas implementaciones, el conjunto de vNRUMI tiene dos longitudes moleculares diferentes. En algunas implementaciones, los vNRUMI tienen seis o siete nucleótidos. En algunas implementaciones, los vNRUMI tienen más de dos longitudes moleculares diferentes, tales como tener tres, cuatro, cinco, seis, siete, ocho, nueve, diez, veinte o más longitudes moleculares diferentes. En algunas implementaciones, las longitudes moleculares se eligen del intervalo 4-100. En algunas implementaciones, las longitudes moleculares se eligen del intervalo 4-20. En algunas implementaciones, las longitudes moleculares se eligen del intervalo de 5 a 15.

En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 10 000 vNRUMI diferentes. En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 1000 vNRUMI diferentes. En algunas implementaciones, el conjunto de vNRUMI incluye no más de aproximadamente 200 vNRUMI diferentes.

En algunas implementaciones, la etapa 134 de la identificación de las lecturas asociadas con el mismo vNRUMI implica obtener, para cada lectura de la pluralidad de lecturas, puntuaciones de alineación con respecto a los vNRUMI. Cada puntuación de alineación indica la similitud entre una subsecuencia de la lectura y un vNRUMI. La subsecuencia se encuentra en una región de la lectura en la que es probable que se ubiquen los nucleótidos derivados del vNRUMI. En otras palabras, en algunas implementaciones, la subsecuencia incluye los primeros nucleótidos de una región donde se espera que esté ubicado el vNRUMI. En algunas implementaciones, el tamaño de la subsecuencia es igual al tamaño del vNRUMI más grande del conjunto de vNRUMI.

En algunas implementaciones, las puntuaciones de alineación se basan en emparejamientos y emparejamientos erróneos/modificaciones de nucleótidos entre la subsecuencia de la lectura y el vNRUMI. En algunas implementaciones, las modificaciones de nucleótidos incluyen sustituciones, adiciones y deleciones de nucleótidos. En algunas implementaciones, la puntuación de alineación penaliza las modificaciones al principio de una secuencia (p. ej., una subsecuencia de una lectura o una secuencia de referencia de un vNRUMI), pero no penaliza las modificaciones al final de la secuencia. La puntuación de alineación refleja la similitud entre la subsecuencia de la lectura y la secuencia de referencia del vNRUMI.

En algunas implementaciones, obtener una puntuación de alineación entre una lectura y un vNRUMI implica: (a) calcular una puntuación de alineación entre el vNRUMI y cada una de las posibles secuencias de prefijo de la subsecuencia de la lectura; (b) calcular una puntuación de alineación entre la subsecuencia de la lectura y cada una de las posibles secuencias de prefijo del vNRUMI; y (c) obtener una puntuación de alineación más alta entre las puntuaciones de alineación calculadas en (a) y (b) como la puntuación de alineación entre la lectura y el vNRUMI.

En algunas implementaciones, la subsecuencia de la lectura tiene una longitud que es igual a la longitud del vNRUMI más largo del conjunto de vNRUMI.

En algunas implementaciones, identificar las lecturas asociadas al mismo vNRUMI incluye seleccionar, para cada lectura de la pluralidad de lecturas, al menos un vNRUMI del conjunto de vNRUMI basándose en las puntuaciones de alineación; y asociar cada lectura de la pluralidad de lecturas con al menos el vNRUMI seleccionado para la lectura. En algunas implementaciones, seleccionar al menos un vNRUMI del conjunto de vNRUMI incluye seleccionar un vNRUMI que tenga la puntuación de alineación más alta de entre el conjunto de vNRUMI.

En algunas implementaciones, se identifica un vNRUMI para obtener la puntuación de alineación más alta. En algunas implementaciones, se identifican dos o más vNRUMI para obtener la puntuación de alineación más alta. En tal caso, la información contextual sobre las lecturas puede usarse para seleccionar uno de los dos o más vNRUMI que deben asociarse con las lecturas para determinar la secuencia de los fragmentos de ADN. Por ejemplo, el número total de lecturas identificadas para un vNRUMI se puede comparar con el número total de lecturas identificadas para otro vNRUMI y un número total superior determina el único vNRUMI que se debe usar para indicar la fuente del fragmento de ADN. En otro ejemplo, la información de secuencia de las lecturas o las ubicaciones de las lecturas en una secuencia de referencia pueden usarse para seleccionar uno de los vNRUMI identificados asociados a las lecturas, utilizándose el vNRUMI seleccionado para determinar la fuente de las lecturas de secuencia.

En algunas implementaciones, se pueden usar dos o más de las puntuaciones de alineación más altas para identificar dos o más vNRUMI para indicar la fuente potencial de cualquier fragmento. La información contextual se puede usar como se mencionó anteriormente para determinar cuál de los vNRUMI indica la fuente real del fragmento de ADN.

La Figura 1E muestra ejemplos de cómo una subsecuencia de una lectura o una secuencia de consulta (Q) puede compararse con dos secuencias de referencia del conjunto de vNRUMI γ = {S1, S2}={AACTTC, CGCTTTCG}. La secuencia de consulta Q incluye los primeros siete nucleótidos de la secuencia de lectura donde se espera que las lecturas se deriven de los vNRUMI.

La secuencia de consulta Q incluye siete nucleótidos GTCTTCG. Q tiene la misma longitud que el vNRUMI más largo del conjunto de vNRUMI γ. La tabla 150 de puntuaciones de alineación muestra las puntuaciones de alineación para las secuencias de prefijo de Q y S1. Por ejemplo, la celda 151 muestra la puntuación de alineación para la secuencia de prefijo de Q (GTCTTC) y la secuencia completa de S1 (AACTTC). La puntuación de alineación tiene en cuenta el número de emparejamientos entre las dos secuencias, así como el número de modificaciones entre las dos secuencias. Por cada nucleótido emparejado, la puntuación aumenta en 1; por cada deleción, adición o sustitución, la puntuación se reduce en 1. Por el contrario, una distancia de Levenshtein es una distancia de modificación, que no tiene en cuenta el número de emparejamientos entre dos secuencias, sino que solo tiene en cuenta el número de adiciones, deleciones y sustituciones.

Al comparar la secuencia de prefijo de Q (GTCTTC) y S1 (AACTTC) nucleótido por nucleótido, hay un emparejamiento erróneo entre G y A, un emparejamiento erróneo entre T y A, un emparejamiento entre C y C, un emparejamiento entre T y T, un emparejamiento entre T y T y un emparejamiento entre C y C. Por lo tanto, la puntuación de alineación para las dos secuencias de prefijo es de 2, como se muestra en la celda 151. La puntuación de alineación no penaliza el extremo de la secuencia Q que tiene un nucleótido G.

En la tabla 150 de puntuaciones de alineación, la columna situada más a la derecha con las puntuaciones de alineación en negrita muestra las puntuaciones de alineación entre todas las posibles subsecuencias de la secuencia de consulta Q y todas las posibles secuencias de prefijo de la secuencia de referencia de vNRUMI S1. La fila inferior de la tabla 150 de puntuaciones de alineación muestra las puntuaciones de alineación entre la secuencia completa S1 y todas las posibles secuencias de prefijo de Q. En diversas implementaciones, la puntuación de alineación más alta de la columna más a la derecha y la fila inferior se selecciona como la puntuación de alineación glocal entre Q y S1. En este ejemplo, la celda 151 tiene el valor más alto, que se determina como la puntuación de alineación glocal entre Q y S1, o g(Q,S1).

La puntuación de alineación más alta de la fila inferior y la columna más a la derecha se utiliza como puntuación de alineación glocal entre dos secuencias. Las diferentes operaciones de cadenas se ponderan por igual en las puntuaciones de alineación ilustradas aquí. Una puntuación de alineación se calcula de la siguiente manera: *n.º de emparejamientos - n.º de inserciones - n.º de deleciones - n.º de sustituciones = n.º de emparejamiento - distancia de Levenshtein.* Sin embargo, como se mencionó anteriormente, en algunas implementaciones, las diferentes operaciones de cadena pueden ponderarse de manera diferente al calcular una puntuación de alineación. Por ejemplo, en algunas implementaciones (no se muestran en la Figura 1E), una puntuación de alineación se puede calcular de la siguiente manera: *n.º de emparejamientos* × *5 - n.º de inserciones* × *4 - n.º de deleciones* × *4 - n.º de sustituciones* × 6 o utilizando otros valores de ponderación.

En las implementaciones descritas anteriormente, las puntuaciones de alineación combinan los efectos de los emparejamientos y las modificaciones de forma lineal, especialmente, mediante adición y/o sustracción. En otras implementaciones, las puntuaciones de alineación pueden combinar los efectos de los emparejamientos y las modificaciones de manera no lineal, tal como mediante operaciones logarítmicas o de multiplicación.

Las puntuaciones de alineación de la columna situada más a la derecha y de la fila inferior indican la similitud entre las secuencias de prefijo, por un lado, y una secuencia completa, por el otro. Cuando el principio de una secuencia de prefijo no coincide con el principio de la secuencia completa, la puntuación de alineación se penaliza. En este sentido, la puntuación de alineación tiene un componente global. Por otro lado, cuando el final de una secuencia de prefijo no coincide con el final de la secuencia completa, la puntuación de alineación de la secuencia no se penaliza. En este sentido, la puntuación de alineación tiene un componente local. Por lo tanto, las puntuaciones de alineación de la columna situada más a la derecha y de la fila inferior pueden describirse como puntuaciones de alineación “glocales”. La puntuación de alineación glocal entre Q y S1 es la puntuación de alineación más alta en la fila más a la derecha y la columna inferior, que es 2, y en la celda 151 para las secuencias de prefijo Q GTCTTC y S1 (AACTTC).

La distancia de Levenshtein entre la secuencia de prefijo Q GTCTTC y S1 (AACTTC) también es 2, porque hay un emparejamiento erróneo entre G y A, un emparejamiento erróneo entre T y A y cuatro emparejamientos para CTTC. Para estas dos secuencias, la distancia de Levenshtein y la puntuación de alineación son las mismas.

En comparación con una puntuación de alineación glocal, una puntuación de alineación global pura requiere la secuencia Q completa por un lado y la secuencia S1 completa por otro lado, que es la puntuación de alineación de la esquina inferior derecha de la tabla 150.

La Tabla 152 de la Figura 1E muestra las puntuaciones de alineación para la secuencia de consulta Q y la secuencia de referencia S2 (CGCTTCG). La puntuación de alineación más alta de la columna situada más a la derecha y de la fila inferior está en la celda 153, que tiene un valor de 4. Es la puntuación de alineación glocal entre Q y S2, o g(Q,S2).

La distancia de Levenshtein entre Q y S2 es idéntica a la distancia de Levenshtein entre Q y S1, porque hay dos emparejamientos erróneos entre las dos secuencias en ambas comparaciones. Sin embargo, g(Q,S2) es mayor que g(Q,S1), porque hay más nucleótidos emparejados entre Q y S2 que entre Q y S1. Especialmente, las puntuaciones de alineación glocal tienen en cuenta no solo las ediciones de nucleótidos (como lo hace la distancia de Levenshtein), sino también los emparejamientos de nucleótidos entre secuencias.

La Figura 1E ilustra que la puntuación de alineación glocal puede proporcionar una mejor corrección de errores que la distancia de Levenshtein o la distancia de modificación, porque la distancia de Levenshtein solo representa el número de modificaciones en la secuencia, mientras que la puntuación de alineación glocal tiene en cuenta tanto el número de modificaciones como el número de emparejamientos entre las secuencias. La Figura 1F proporciona un ejemplo que ilustra que la puntuación de alineación glocal puede proporcionar una mejor supresión de errores que la puntuación de alineación global, porque la puntuación de alineación glocal no penaliza en exceso los emparejamientos erróneos debidos a la inserción, deleción o sustitución al final de la secuencia.

El ejemplo de la Figura 1F usa un conjunto diferente de secuencias de vNRUMI, γ = {S1,S2} = {TTGTGAC,GGCCAT}. En el proceso de procesamiento de muestras S1 se usa para etiquetar una molécula de ADN. La secuencia de esta molécula es m_0 = ***TTGTGACTNNNNN*** (SEQ ID NO: 1). Durante la secuenciación, se produce un único error de inserción y la secuencia GCA se inserta en m_0, creando m_1 = ***TTGGCATGACTNNNNN*** (SEQ ID NO: 2). Para corregir este error y recuperar el UMI apropiado para esta secuencia, un proceso toma los primeros 7 pares de bases como la secuencia de consulta, **Q** = ***TTGGCAT.*** El proceso compara ***Q*** con cada secuencia de γ.

Se obtiene una tabla 160 de puntuaciones de alineación para **g(Q, S1)** y se muestra en la Figura 1F. Y similarmente, se obtiene una tabla 163 de puntuaciones de alineación para **g(Q, S2).**

Si se usa un esquema de alineación global en lugar de una puntuación de alineación glocal, se usaría la puntuación de la esquina inferior derecha de las celdas 161 y 164, que tienen un valor de 2 en ambos casos. Una alineación óptima de Q (TTGGCAT) y S1 (TTGTGAC) consiste en alinear TTG-GCAT con TTGTG-AC, donde los guiones representan inserciones o huecos. Esta alineación implica 5 emparejamientos, 2 inserciones y 1 sustitución, lo que proporciona una puntuación de alineación de 5-2-1 = 2. Una alineación óptima de Q (TTGGCAT) y S2 (GGCCAT) consiste en alinear TTGGC-AT y --GGCCAT. Esta alineación implica 5 emparejamientos y 3 inserciones, lo que proporciona una puntuación de alineación de 5-3 = 2. Al utilizar una puntuación de alineación global, no se puede determinar de manera concluyente cuál de los S1 y S2 tiene más probabilidades de ser el vNRUMI real.

Sin embargo, al usar un esquema de alineación glocal, que usa el valor máximo de la última fila y columna, el proceso obtiene una puntuación de alineación de 3 para las secuencias de prefijo TTGGC y S1 (TTGTGAC) de Q, que se convierte en la puntuación glocal de S1 y es superior a la puntuación glocal de S2 (2). Como tal, el proceso puede asociar correctamente Q con S1.

Volviendo a la Figura 1C, la etapa 135 implica determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI. En algunas implementaciones, la determinación de la secuencia del fragmento de ADN implica agrupar las lecturas asociadas con el mismo vNRUMI para obtener una secuencia de consenso, lo que se puede lograr como se describe además a continuación en la memoria. En algunas implementaciones, la secuencia de consenso se basa en las puntuaciones de calidad de las lecturas, así como en la secuencia de las lecturas. De forma adicional o alternativamente, se puede usar otra información contextual, tal como la posición de las lecturas, para determinar la secuencia de consenso.

En algunas implementaciones, la determinación de la secuencia del fragmento de ADN también implica identificar las lecturas que tienen la misma posición o posiciones similares en una secuencia de referencia. El método determina después la secuencia del fragmento de ADN utilizando lecturas que están asociadas con el mismo vNRUMI y tienen la misma posición o posiciones similares en la secuencia de referencia.

En algunas implementaciones, determinar la secuencia del fragmento de ADN implica identificar, entre las lecturas asociadas con el mismo vNRUMI, las lecturas que comparten un UMI virtual común o UMI virtuales similares, donde los UMI virtuales comunes se encuentran en el fragmento de ADN. El método también implica determinar la secuencia del fragmento de ADN utilizando únicamente lecturas que estén ambas asociadas con el mismo vNRUMI y compartan los mismos UMI virtuales o UMI virtuales celulares.

En algunas implementaciones, los adaptadores de secuenciación que tienen vNRUMI pueden prepararse mediante un proceso representado en la Figura 1D y descrito además a continuación en la memoria.

Diseño de UMI

UMI físicos

En algunas implementaciones de los adaptadores descritos anteriormente, los UMI físicos de los adaptadores incluyen UMI aleatorios. En algunas implementaciones, cada UMI aleatorio es diferente de cualquier otro UMI aleatorio aplicado

a fragmentos de ADN. En otras palabras, los UMI aleatorios se seleccionan aleatoriamente sin reemplazo entre un conjunto de UMI que incluye todos los UMI diferentes posibles dada(s) la(s) longitud(es) de la secuencia. En otras implementaciones, los UMI aleatorios se seleccionan aleatoriamente con reemplazo. En estas implementaciones, dos adaptadores pueden tener el mismo UMI debido al azar.

En algunas implementaciones, los UMI físicos utilizados en un proceso son un conjunto de NRUMI que se seleccionan de entre un grupo de secuencias candidatas utilizando un planteamiento ambicioso que maximiza las diferencias entre los UMI seleccionados, como se describe además a continuación en la memoria. En algunas implementaciones, los NRUMI tienen longitudes moleculares variables o heterogéneas, formando un conjunto de vNRUMI. En algunas implementaciones, el grupo de secuencias candidatas se filtra para eliminar determinadas secuencias antes de proporcionarlo para seleccionar un conjunto de UMI usados en una reacción o proceso.

Los UMI aleatorios proporcionan un mayor número de UMI únicos que los UMI no aleatorios de la misma longitud de secuencia. En otras palabras, los UMI aleatorios tienen más probabilidades de ser únicos que los UMI no aleatorios. Sin embargo, en algunas implementaciones, los UMI no aleatorios pueden ser más fáciles de fabricar o tienen una mayor eficiencia de conversión. Cuando los UMI no aleatorios se combinan con otra información, tal como la posición de la secuencia y el UMI virtual, pueden proporcionar un mecanismo eficaz para indexar las moléculas fuente de los fragmentos de ADN.

Estructura de vNRUMI

En algunas implementaciones, los adaptadores de secuenciación que tienen vNRUMI se pueden preparar mediante un planteamiento ambicioso representado en la Figura 1D. El proceso implica (a) proporcionar un conjunto de secuencias de oligonucleótidos que tienen dos longitudes moleculares diferentes; y (b) seleccionar un subconjunto de secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos, todas las distancias de modificación entre las secuencias de oligonucleótidos del subconjunto que cumplen un valor umbral. El subconjunto de secuencias de oligonucleótidos forma un conjunto de vNRUMI. El método también implica (c) sintetizar una pluralidad de adaptadores de secuenciación, el adaptador de secuenciación tiene una región hibridada bicatenaria, un extremo 5' monocatenario, un extremo 3' monocatenario como se muestra en las Figuras 2A, y al menos un vNRUMI del conjunto de vNRUMI.

La Figura 1D ilustra un proceso 140 para crear adaptadores de secuenciación que tengan vNRUMI. El proceso 140 comienza proporcionando un conjunto de secuencias de oligonucleótidos (β) que tienen al menos dos longitudes moleculares diferentes. Véase el bloque 141.

En diversas implementaciones, los UMI no aleatorios se preparan considerando diversos factores, incluidos aunque no de forma limitativa, los medios para detectar errores dentro de las secuencias de UMI, la eficiencia de conversión, la compatibilidad de los ensayos, el contenido de GC, los homopolímeros y las consideraciones de fabricación.

En algunas implementaciones, antes de la operación 141, algunas de las secuencias de oligonucleótidos se eliminan del conjunto completo de todas las permutaciones posibles de nucleótidos dadas las longitudes moleculares específicas del conjunto de vNRUMI. Por ejemplo, si los vNRUMI tienen longitudes moleculares de seis y siete nucleótidos, todas las permutaciones de secuencias posibles incluyen un grupo completo de $4^6 + 4^7 = 20\,480$ secuencias. Determinadas secuencias de oligonucleótidos se eliminan del grupo para proporcionar el conjunto de secuencias de oligonucleótidos β.

En algunas implementaciones, las secuencias de oligonucleótidos que tienen tres o más bases idénticas consecutivas se eliminan del grupo para proporcionar el conjunto β. En algunas implementaciones, se eliminan las secuencias de oligonucleótidos que tienen un número combinado de bases de guanina y citosina (G y C) de menos de dos. En algunas implementaciones, se eliminan las secuencias de oligonucleótidos que tienen un número combinado de bases de guanina y citosina de más de cuatro. En algunas implementaciones, se eliminan las secuencias de oligonucleótidos que tienen la misma base en las dos últimas posiciones de la secuencia. La secuencia comienza desde el extremo opuesto al extremo unido a los fragmentos de ADN.

En algunas implementaciones, se eliminan las secuencias de oligonucleótidos que tienen una subsecuencia que se empareja con el extremo 3' de cualquier cebador de secuenciación.

En algunas implementaciones, se eliminan las secuencias de oligonucleótidos que tienen una base de timina (T) en la última posición de las secuencias de nucleótidos. Un vNRUMI unido a un extremo A de un fragmento de ácido nucleico procesado resultará en una subsecuencia de una lectura que tenga la secuencia de vNRUMI y una base T hibridada al final de la secuencia de vNRUMI, siendo la T complementaria de la base A en el extremo A. Filtrar las secuencias candidatas que tienen una base T en la última posición evita la confusión entre tales secuencias candidatas y la subsecuencia de lecturas derivadas de cualquier vNRUMI.

El proceso 140 procede seleccionando una secuencia de oligonucleótidos (S_0) de β. Véase el bloque 142. En algunas implementaciones, S_0 puede elegirse al azar del conjunto de secuencias de oligonucleótidos.

El proceso 140 implica además añadir S_0 a un conjunto en expansión γ de secuencias de oligonucleótidos y eliminar S_0 del conjunto β. Véase el bloque 143.

El proceso 140 implica además seleccionar la secuencia de oligonucleótidos S_i de β, S_i maximizando la función de distancia $d(S_i, \gamma)$, que es una distancia de modificación mínima entre S_i y cualquier secuencia de oligonucleótidos del conjunto γ. Véase el bloque 144. En algunas implementaciones, la distancia de edición es la distancia de Levenshtein.

En algunas implementaciones, cuando la secuencia es más corta que la longitud máxima de los vNRUMI, se incorporan una o más bases en el extremo de la secuencia al calcular la distancia de Levenshtein o la distancia de modificación. En algunas implementaciones, si la secuencia es una base más corta que la longitud máxima de los vNRUMI, se añade una base de timina (T) al extremo de la secuencia. Esta base T se añade para reflejar un saliente de base T al extremo de un adaptador complementario a la base A al extremo de un fragmento de ADN que ha sido sometido a un procesamiento de dA-tailing como se describe en cualquier otro sitio de la presente memoria. En algunas implementaciones, si la secuencia es de más de una base más corta que la longitud máxima de los vNRUMI, se añade una base T al extremo de la secuencia, y después se añaden una o más bases aleatorias después de la base T para crear una secuencia que tenga una longitud molecular igual a la longitud máxima de los vNRUMI. En otras palabras, se pueden incorporar múltiples combinaciones diferentes de bases aleatorias después de la base T para crear secuencias que abarquen todas las posibles secuencias observadas. Por ejemplo, si los vNRUMI tienen longitudes de 6 y 8, se pueden obtener cuatro derivaciones de un 6mero incorporando TA, TC, TG y TT.

El proceso 140 procede a determinar si la función de distancia $d(S_i, \gamma)$ cumple el valor umbral. En algunas implementaciones, el valor umbral puede requerir que la función de distancia (p. ej., una distancia de Levenshtein rellena) sea de al menos 3. Si la función de distancia $d(S_i, \gamma)$ alcanza el umbral, el proceso procede a añadir S_i al conjunto en expansión gamma γ y elimina S_i del conjunto β. Véase la rama "Sí" de la decisión 145 y el bloque 146. Si la función de distancia no alcanza el valor umbral, el proceso 140 no añade S_i al conjunto en expansión γ, y el proceso procede a sintetizar la pluralidad de adaptadores de secuenciación, donde cada adaptador de secuenciación tiene al menos un vNRUMI en el conjunto en expansión γ. Véase la rama sin decisión de 145 que apunta al bloque 148.

Después de la etapa 146, el proceso 140 implica además una operación de decisión sobre si es necesario considerar más secuencias del conjunto β. Si es así, el proceso vuelve al bloque 144 para seleccionar más secuencias de oligonucleótidos del conjunto β que maximiza la función de distancia. Se pueden considerar diversos factores para determinar si es necesario examinar además más secuencias del conjunto β. Por ejemplo, en algunas implementaciones, cuando se ha obtenido el número deseado de secuencias, el proceso ya no necesita considerar más secuencias de los datos del conjunto de secuencias.

Cuando se decide que no es necesario considerar más secuencias, el proceso 140 procede a sintetizar la pluralidad de adaptadores de secuenciación, donde cada adaptador tiene al menos un vNRUMI en el conjunto de secuencias γ. Véase la rama sin decisión de la operación 147 que apunta a la operación 148. En algunas implementaciones, cada adaptador de secuenciación tiene el vNRUMI en una cadena de los adaptadores de secuenciación. En algunas implementaciones, los adaptadores de secuenciación que tienen cualquiera de las formas ilustradas en la Figura 2A se sintetizan en la operación 148. En algunas implementaciones, cada adaptador de secuenciación tiene solo un vNRUMI. En algunas implementaciones, cada adaptador tiene un vNRUMI en cada cadena de los adaptadores de secuenciación. En algunas implementaciones, cada adaptador de secuenciación tiene un vNRUMI en cada cadena del adaptador de secuenciación en la región hibridada bicatenaria.

En algunas implementaciones, el proceso puede implementarse mediante el pseudocódigo siguiente.

```
algoritmo vNRUMI_dist:
    entrada: conjunto S de secuencias de vNRUMI, secuencia de consulta Q
    salida: entero d que representa la distancia de Q a S
    dejar que las distancias sean una lista de todas las distancias encontradas
    para cada secuencia s en S:
        si la longitud(s) es inferior a la longitud máxima de cualquier secuencia en S:
            añadir una "T" a s
        si la longitud(Q) es inferior a la longitud máxima de cualquier secuencia en S:
            añadir una "T" a Q
        añadir Levenshtein (s, Q) a las distancias
        devolver el valor mínimo en distancias

algoritmo generate_vNRUMI_set:
    entrada: conjunto X que contiene secuencias de vNRUMI potenciales/candidatas
             entero N que indica el número de vNRUMI deseados en el conjunto
    salida: conjunto Y que contiene un conjunto de como máximo N vNRUMI
    elegir un elemento aleatorio de X, añadirlo a Y, eliminarlo de X
    mientras que el número de secuencias en Y < N:
```

```
            almacenar vNRUMI_dist para cada candidato en X contra Y
            Z = vNRUMI_dist máxima encontrada
            Si Z >= 3:
                S = conjunto de todas las secuencias que tienen un vNRUMI_dist de Z
                Selegido = elegir un objeto aleatorio de S, preferiblemente más corto
    secuencias
                añadir Selegido Y, eliminarlo de X
            si no:
                devolver Y
        devolver Y
```

A continuación, se presenta un ejemplo ficticio para ilustrar cómo se pueden obtener vNRUMI de acuerdo con el proceso y el algoritmo descritos anteriormente. El ejemplo ficticio muestra cómo se pueden producir vNRUMI a partir de un grupo de cinco secuencias candidatas, que después se utilizan para mapear las lecturas de secuencias observadas. Tenga en cuenta que, dado que este es un ejemplo ficticio sobre un espacio de secuencia significativamente más pequeño que el que usaríamos/encontraríamos en la práctica, no se pueden abordar todos los aspectos de las características de los vNRUMI.

En este ejemplo ficticio, el proceso tiene como objetivo construir un conjunto de 3 secuencias de vNRUMI partiendo de un conjunto de 6meros y 7meros (pero resultó en solo 2 secuencias de vNRUMI). Para simplificar, presuponga que todo el espacio de posibles 6 meros y 7 meros consiste en las siguientes 5 secuencias:

AACTTC

AACTTCA

AGCTTCG

CGCTTCG

CGCTTC

Tenga en cuenta que se presupone que todas estas 5 secuencias han pasado todos los filtros bioquímicos que están implementados. A un nivel muy alto, este algoritmo subdivide el grupo de secuencias de entrada al tiempo que maximiza una distancia de modificación (una distancia de Levenshtein) entre las secuencias elegidas. Lo hace utilizando un planteamiento ambicioso: en cada variación, elige una secuencia que maximiza la función de distancia. La función de distancia, en este caso, es la distancia mínima de modificación entre la secuencia que se va a añadir y cualquier secuencia que ya esté en el conjunto. Esto se puede expresar matemáticamente de la siguiente manera:

$$d(s,\gamma) = \min\left(levenshtein(s,x)\ \forall\ x \in \gamma\right)$$

En el ejemplo siguiente, el conjunto de vNRUMI (n=3) que se está construyendo se denotará como γ, y el conjunto de secuencias candidatas de entrada se denotará como β.

$$\gamma = \{\ \ \},\beta = \{AACTTC, AACTTCA, AGCTTCG, CGCTTCG, CGCTTC\}$$

Como no hay secuencias en γ, la función de distancia d no está definida para cada una de las 5 secuencias. En caso de empate para la mejor elección, siempre elegimos al azar una de las candidatas empatadas, prefiriendo las secuencias más cortas. Aquí, el ejemplo selecciona la secuencia de 6meros AACTTC. Añade la secuencia a γ y la elimina del grupo de secuencias candidatas.

$$\gamma = \{AACTTC\},\beta = \{AACTTCA, AGCTTCG, CGCTTCG, CGCTTC\}$$

Se calcula la métrica de distancia $d(s, \gamma)\forall s \in \beta$.

$d(AACTTCA,\gamma)$ = 1, ya que solo se necesita una edición (adición de una A) para pasar del elemento único de γ a AACTTCA y, por lo tanto, la función de distancia es 1.

$d(AGCTTCG,\gamma)$ = 2, ya que se necesitan dos modificaciones para pasar de esta secuencia a la secuencia que ya está en γ.

$d(CGCTTCG,\gamma)$ = 3, ya que se necesitan tres modificaciones para pasar de esta secuencia a la secuencia que ya está en γ.

$d(CGCTTC,\gamma) = 2$, dado que la secuencia en comparación es un hexámero, en algunas implementaciones, se añade una base "T" en el extremo de la misma para simular el proceso de hibridación, en el que una base T complementaria al extremo "A" se hibrida a la secuencia adaptadora. La justificación es que cuando los practicantes traten de identificar el NRUMI más adelante, considerarán tanto el primer hexámero como el primer heptámero. Al añadir esta base T, se garantiza que, al observar el heptámero, siga sin estar demasiado cerca de ningún otro NRUMI. Al comparar CGCTTCT con AACTTC, se requieren dos modificaciones.

Dado que la función de distancia máxima es 3, producida por la secuencia CGCTTCG, y esta distancia supera nuestro umbral mínimo (de 3), el proceso añade CGCTTCG a γ y lo elimina de β.

$$\gamma = \{AACTTC, CGCTTCG\}, \beta = \{AACTTCA, AGCTTCG, CGCTTC\}$$

A continuación, el proceso procede a calcular la métrica de distancia $d(s,\gamma) \forall s \in \beta$, ya que hay menos del número deseado (3) de secuencias en el conjunto de vNRUMI.

$d(AACTTCA,\gamma) = 1$. Como se calculó en la etapa anterior, la distancia de modificación entre esta secuencia y la primera secuencia de vNRUMI, $s_1 = AACTTC$, es 1. La distancia de modificación entre esta secuencia y la segunda secuencia de vNRUMI, $s_2 = CGCTTCG$, es 3. La función de distancia toma el mínimo de todas las distancias de modificación entre la secuencia de consulta y cualquier secuencia existente, y $\min(3,1) = 1$, por lo que la función de distancia es 1.

$d(AGCTTCG, \gamma) = 1$. Como se calculó en la etapa anterior, la distancia de modificación entre esta secuencia y s_1 es 2. La distancia de modificación entre esta secuencia y s_2 es 1. Por lo tanto, la función de distancia es la menor de 2 y 1 (que es 1).

$d(CGCTTC,\gamma) = 1$. Como antes, el proceso incorpora una T a esta secuencia para convertirla en CGCTTCT. La distancia entre la consulta alargada y s_1 es 2, como se determinó anteriormente. La distancia entre la consulta alargada y s_2 es 1, por lo que la función de distancia es 1.

Después de calcular todas las funciones de distancia para todas las secuencias candidatas, ninguna de ellas satisface nuestro requisito invariable de una distancia de modificación de al menos 3. Este requisito hace muy improbable que mutaciones aleatorias muten una secuencia de vNRUMI en algo parecido a otra. Por lo tanto, devolvemos este conjunto de 2 secuencias de vNRUMI, $\gamma = \{AACTTC, CGCTTCG\}$. Cabe señalar que las dos secuencias de vNRUMI son las mismas que S1 y S2 de la Figura 1E descritas anteriormente, y podrían asociarse con lecturas para determinar el segmento fuente de las lecturas como se describe con referencia a la Figura 1E.

UMI virtuales

En cuanto a los UMI virtuales, los UMI virtuales que se definen en, o con respecto a, las posiciones terminales de moléculas de ADN fuente, pueden definir de manera única o casi única moléculas de ADN fuente individuales cuando las ubicaciones de las posiciones terminales son generalmente aleatorias, como ocurre con algunos procedimientos de fragmentación y con el ADNIc de origen natural. Cuando la muestra contiene relativamente pocas moléculas de ADN fuente, los UMI virtuales pueden identificar por sí mismos moléculas individuales de ADN fuente. El uso de una combinación de dos UMI virtuales, cada uno asociado a un extremo diferente de una molécula de ADN fuente, aumenta la probabilidad de que los UMI virtuales por sí solos puedan identificar de forma exclusiva las moléculas de ADN fuente. Por supuesto, incluso en situaciones donde uno o dos UMI virtuales no pueden por sí solos identificar de forma única moléculas de ADN fuente, puede tener éxito la combinación de tales UMI virtuales con uno o más UMI físicos.

Si dos lecturas proceden del mismo fragmento de ADN, dos subsecuencias que tengan los mismos pares de bases también tendrán la misma ubicación relativa en las lecturas. Por el contrario, si dos lecturas proceden de dos fragmentos de ADN diferentes, es poco probable que dos secuencias con los mismos pares de bases tengan exactamente la misma ubicación relativa en las lecturas. Por lo tanto, si dos o más secuencias de dos o más lecturas tienen los mismos pares de bases y la misma ubicación relativa en las dos o más lecturas, se puede inferir que las dos o más lecturas proceden del mismo fragmento.

En algunas implementaciones, las subsecuencias en o cerca de los extremos de un fragmento de ADN se utilizan como UMI virtuales. Esta elección de diseño tiene algunas ventajas prácticas. En primer lugar, las ubicaciones relativas de estas subsecuencias en las lecturas son fáciles de determinar, ya que se encuentran al principio o cerca del principio de las lecturas y el sistema no necesita utilizar un desplazamiento para encontrar el UMI virtual. Además, como primero se secuencian los pares de bases de los extremos de los fragmentos, esos pares de bases están disponibles aunque las lecturas sean relativamente cortas. Además, los pares de bases que se determinan antes en una lectura larga tienen una tasa de error de secuenciación menor que los que se determinan después. En otras implementaciones, sin embargo, las subsecuencias situadas lejos de los extremos de las lecturas pueden utilizarse como UMI virtuales, pero puede ser necesario determinar sus posiciones relativas en las lecturas para inferir que las lecturas se obtienen del mismo fragmento.

Una o varias subsecuencias de una lectura pueden utilizarse como UMI virtuales. En algunas implementaciones, dos subsecuencias, cada una rastreada desde un extremo diferente de la molécula de ADN fuente, se utilizan como UMI virtuales. En diversas implementaciones, los UMI virtuales son de aproximadamente 24 pares de bases o más cortos, aproximadamente 20 pares de bases o más cortos, aproximadamente 15 pares de bases o más cortos, aproximadamente 10 pares de bases o más cortos, aproximadamente 9 pares de bases o más cortos, aproximadamente 8 pares de bases o más cortos, aproximadamente 7 pares de bases o más cortos o aproximadamente 6 pares de bases o más cortos. En algunas implementaciones, los UMI virtuales son de aproximadamente 6 a 10 pares de bases. En otras implementaciones, los UMI virtuales son de aproximadamente 6 a 24 pares de bases.

Adaptadores

Además del diseño de adaptador descrito en el ejemplo de flujo de trabajo 100 con referencia a la Figura 1A anterior, se pueden utilizar otros diseños de adaptadores en diversas implementaciones de los métodos y sistemas expuestos en la presente memoria. La Figura 2A ilustra esquemáticamente cinco diseños diferentes de adaptador con uno o más UMI que pueden adoptarse en las diversas implementaciones.

La Figura 2A(i) muestra un adaptador de doble índice estándar de TruSeq® de Illumina. El adaptador es parcialmente bicatenario y se forma hibridando dos oligonucleótidos correspondientes a las dos cadenas. Las dos cadenas tienen un número de pares de bases complementarias (p. ej. 12–17 pb), lo que permite hibridar los dos oligonucleótidos del extremo con un fragmento de ADNbc. Un fragmento de ADNbc que se va a ligar a ambos extremos para la lectura en ambos extremos también se denomina inserto. Otros pares de bases no son complementarios en las dos cadenas, lo que da lugar a un adaptador en forma de horquilla con dos salientes flexibles. En el ejemplo de la Figura 2A(i), los pares de bases complementarios forman parte de la secuencia del cebador de lectura 2 y de la secuencia del cebador de lectura 1. Cadena abajo de la secuencia del cebador de lectura 2 hay un saliente de un solo nucleótido 3'-T, que proporciona un saliente complementario al saliente de un solo nucleótido 3'-A de un fragmento de ADNbc que se va a secuenciar, lo que puede facilitar la hibridación de los dos salientes. La secuencia del cebador de lectura 1 se encuentra en el extremo 5' de la cadena complementaria, a la que se une un grupo fosfato. El grupo fosfato es necesario para ligar el extremo 5' de la secuencia de cebador 1 de lectura al saliente 3'-A del fragmento de ADN. En la cadena que tiene el saliente 5' (la cadena superior), en dirección de 5' a 3', el adaptador tiene una secuencia P5, una secuencia de índice i5 y la secuencia de cebador de lectura 2. En la cadena que tiene el saliente 3', en dirección de 3' a 5', el adaptador tiene una secuencia P7', una secuencia de índice i7 y la secuencia del cebador de lectura 1. Los oligonucleótidos P5 y P7' son complementarios a los cebadores de amplificación unidos a la superficie de las cubetas de lectura de una plataforma de secuenciación de Illumina. En algunas implementaciones, las secuencias de índice proporcionan medios para realizar un seguimiento de la fuente de una muestra, permitiendo así la multiplexación de múltiples muestras en la plataforma de secuenciación.

La Figura 2A(ii) muestra un adaptador con un único UMI físico que sustituye a la región de índice i7 del adaptador de doble índice estándar mostrado en la Figura 2A(i). Este diseño del adaptador refleja el mostrado en el ejemplo de flujo de trabajo descrito anteriormente en relación con la Figura 1B. En ciertas realizaciones, los UMI físicos α y β están diseñados para estar solo en el brazo 5' de los adaptadores bicatenarios, dando lugar a productos de ligamiento que tienen solo un UMI físico en cada cadena. En comparación, los UMI físicos incorporados en ambas cadenas de los adaptadores dan lugar a productos de ligamiento que tienen dos UMI físicos en cada cadena, lo que duplica el tiempo y el coste de secuenciar los UMI físicos. Sin embargo, la presente exposición incorpora métodos que emplean UMI físicos en ambas cadenas de los adaptadores, como se representa en las Figuras 2A(iii)-2A(vi), que proporcionan información adicional que puede utilizarse para agrupar diferentes lecturas con el fin de obtener secuencias de consenso.

En algunas implementaciones, los UMI físicos de los adaptadores incluyen UMI aleatorios. En algunas implementaciones, los UMI físicos de los adaptadores incluyen UMI no aleatorios.

La Figura 2A(iii) muestra un adaptador que tiene dos UMI físicos añadidos al adaptador de doble índice estándar. Los UMI físicos mostrados aquí pueden ser UMI aleatorios o UMI no aleatorios. El primer UMI físico está cadena arriba de la secuencia de índice i7, y el segundo UMI físico está cadena arriba de la secuencia de índice i5. La Figura 2A(iv) muestra un adaptador que también tiene dos UMI físicos añadidos al adaptador de doble índice estándar. El primer UMI físico se encuentra cadena abajo de la secuencia de índice i7, y el segundo UMI físico se encuentra cadena abajo de la secuencia de índice i5. Del mismo modo, los dos UMI físicos pueden ser UMI aleatorios o no aleatorios.

Un adaptador que tenga dos UMI físicos en los dos brazos de la región monocatenaria, tal como los mostrados en 2A(iii) y 2A(iv), puede unir dos cadenas de un fragmento de ADN bicatenario, si se conoce *a priori* o *a posteriori* la información que asocia los dos UMI físicos no complementarios. Por ejemplo, un investigador puede conocer las secuencias del UMI 1 y el UMI 2 antes de integrarlos al mismo adaptador en el diseño mostrado en la Figura 2A(iv). Esta información de asociación puede utilizarse para inferir que las lecturas que tienen el UMI 1 y el UMI 2 proceden de dos cadenas del fragmento de ADN al que se ligó el adaptador. Por lo tanto, se pueden agrupar no solo lecturas que tengan el mismo UMI físico, sino también lecturas que tengan cualquiera de los dos UMI físicos no

complementarios. Curiosamente, y como se comenta más adelante, un fenómeno denominado "salto de UMI" puede complicar la inferencia de asociación entre UMI físicos en regiones monocatenarias de adaptadores.

Los dos UMI físicos en las dos cadenas de los adaptadores de la Figura 2A(iii) y la Figura 2A(iv) no están situadas en el mismo sitio ni son complementarios entre sí. Sin embargo, la presente exposición incorpora métodos que emplean UMI físicos que se encuentran en el mismo sitio en dos cadenas del adaptador y/o complementarios entre sí. La Figura 2A(v) muestra un adaptador dúplex en el que los dos UMI físicos son complementarios en una región bicatenaria en o cerca del extremo del adaptador. Los dos UMI físicos pueden ser UMI aleatorios o no aleatorios. La Figura 2A(vi) muestra un adaptador similar pero más corto que el de la Figura 2A(v), pero no incluye las secuencias de índice ni las secuencias P5 y P7' complementarias a los cebadores de amplificación de la superficie de la cubeta de lectura. Del mismo modo, los dos UMI físicos pueden ser UMI aleatorios o no aleatorios.

En comparación con los adaptadores que tienen uno o más UMI físicos monocatenarios en brazos monocatenarios, los adaptadores que tienen un UMI físico bicatenario en la región bicatenaria pueden proporcionar un enlace directo entre dos cadenas de un fragmento de ADN bicatenario al que se liga el adaptador, como se muestra en la Figura 2A(v) y en la Figura 2A(vi). Dado que las dos cadenas de un UMI físico bicatenario son complementarias entre sí, la asociación entre las dos cadenas del UMI bicatenario se refleja intrínsecamente en las secuencias complementarias, y puede establecerse sin necesidad de información *a priori* ni *a posteriori*. Esta información puede utilizarse para deducir que las lecturas que tienen las dos secuencias complementarias de un UMI físico bicatenario de un adaptador proceden del mismo fragmento de ADN al que se ligó el adaptador, pero las dos secuencias complementarias del UMI físico están ligadas al extremo 3' en una cadena y al extremo 5' en la otra cadena del fragmento de ADN. Por lo tanto, se pueden agrupar no solo lecturas que tengan el mismo orden de dos secuencias de UMI físicos en dos extremos, sino también lecturas que tengan el orden inverso de dos secuencias complementarias en dos extremos.

En algunas realizaciones, puede ser ventajoso emplear UMI físicos relativamente cortos porque los UMI físicos cortos son más fáciles de incorporar a los adaptadores. Además, los UMI físicos más cortos son más rápidos y fáciles de secuenciar en los fragmentos amplificados. Sin embargo, a medida que los UMI físicos se hacen muy cortos, el número total de UMI físicos diferentes puede llegar a ser inferior al número de moléculas adaptadoras necesarias para el procesamiento de la muestra. Para disponer de suficientes adaptadores, habría que repetir el mismo UMI en dos o más moléculas de adaptador. En tal escenario, los adaptadores que tengan los mismos UMI físicos pueden ligarse a múltiples moléculas de ADN fuente. Sin embargo, estos UMI físicos cortos pueden proporcionar suficiente información, cuando se combinan con otra información como los UMI virtuales y/o las ubicaciones de alineación de las lecturas, para identificar de forma exclusiva las lecturas como derivadas de un polinucleótido o fragmento de ADN fuente concreto de una muestra. Esto es así porque, aunque puede ligarse el mismo UMI físico a dos fragmentos diferentes, es poco probable que los dos fragmentos diferentes tengan también los mismos lugares de alineación, o subsecuencias coincidentes que sirvan como UMI virtuales. Por tanto, si dos lecturas tienen el mismo UMI físico corto y la misma ubicación de alineación (o el mismo UMI virtual), es probable que las dos lecturas procedan del mismo fragmento de ADN.

Además, en algunas implementaciones, el agrupamiento de lectura se basa en dos UMI físicos en los dos extremos de un inserto. En tales implementaciones, se combinan dos UMI físicos muy cortos (p. ej., 4 pb) para determinar la fuente de los fragmentos de ADN, proporcionando la longitud combinada de los dos UMI físicos información suficiente para distinguirlos entre diferentes fragmentos.

En diversas implementaciones, los UMI son de aproximadamente 12 pares de bases o más cortos, aproximadamente 11 pares de bases o más cortos, aproximadamente 10 pares de bases o más cortos, aproximadamente 9 pares de bases o más cortos, aproximadamente 8 pares de bases o más cortos, aproximadamente 7 pares de bases o más cortos, aproximadamente 6 pares de bases o más cortos, aproximadamente 5 pares de bases o más cortos, aproximadamente 4 pares de bases o más cortos o aproximadamente 3 pares de bases o más cortos. En algunas implementaciones donde los UMI físicos son UMI no aleatorios, los UMI son de aproximadamente 12 pares de bases o más cortos, aproximadamente 11 pares de bases o más cortos, aproximadamente 10 pares de bases o más cortos, aproximadamente 9 pares de bases o más cortos, aproximadamente 8 pares de bases o más cortos, aproximadamente 7 pares de bases o más cortos o aproximadamente 6 pares de bases o más cortos.

El salto de UMI puede afectar a la inferencia de asociación entre UMI físicos en uno o ambos brazos de los adaptadores, tales como en los adaptadores de las Figuras 2A(ii)-(iv). Se ha observado que al aplicar estos adaptadores a fragmentos de ADN, los productos de amplificación pueden incluir un mayor número de fragmentos con UMI físicos únicos que el número real de fragmentos de la muestra.

Además, cuando se aplican adaptadores que tienen UMI físicos en ambos brazos, se supone que los fragmentos amplificados que tienen un UMI físico común en un extremo tienen otro UMI físico común en otro extremo. Sin embargo, en ocasiones este no es el caso. Por ejemplo, en el producto de una reacción de amplificación, algunos fragmentos pueden tener un primer UMI físico y un segundo UMI físico en sus dos extremos; otros fragmentos pueden tener el segundo UMI físico y un tercer UMI físico; otros fragmentos pueden tener el primer UMI físico y el tercer UMI físico; otros fragmentos pueden tener el tercer UMI físico y un cuarto UMI físico, y así sucesivamente. En este ejemplo,

puede ser difícil determinar el fragmento o fragmentos de origen de estos fragmentos amplificados. Al parecer, durante el proceso de amplificación, el UMI físico puede haberse "intercambiado" por otro UMI físico.

Un posible planteamiento para abordar este problema de salto de UMI considera que solo los fragmentos que comparten ambos UMI proceden de la misma molécula de origen, mientras que los fragmentos que solo comparten un UMI se excluirán del análisis. Sin embargo, algunos de estos fragmentos que solo comparten un UMI físico pueden proceder de la misma molécula que los que comparten ambos UMI físicos. Si se excluyen de la consideración los fragmentos que comparten un solo UMI físico, puede perderse información útil. Otro posible planteamiento considera que cualquier fragmento que tenga un UMI físico común procede de la misma molécula de origen. Pero este planteamiento no permite combinar dos UMI físicos en dos extremos de los fragmentos para su análisis posterior. Además, en cualquiera de los dos planteamientos, para el ejemplo anterior, no se consideraría que los fragmentos que comparten el primer y el segundo UMI físico proceden de la misma molécula fuente que los fragmentos que comparten el tercer y el cuarto UMI físico. Esto puede ser cierto o no. Un tercer planteamiento puede abordar el problema del salto de UMI utilizando adaptadores con UMI físicos en ambas cadenas de la región monocatenaria, tales como los adaptadores de las Figuras 2A(v)-(vi). Además, se explica a continuación una descripción de un mecanismo hipotético subyacente al salto de UMI.

La Figura 2B ilustra un proceso hipotético en el que se produce un salto de UMI en una reacción de PCR en la que intervienen adaptadores que tienen UMI físicos en ambas cadenas de la región bicatenaria. Los dos UMI físicos pueden ser UMI aleatorios o no aleatorios. El mecanismo subyacente real de salto de UMI y el proceso hipotético descrito aquí no afectan a la utilidad de los adaptadores y métodos expuestos en la presente memoria. La reacción PCR comienza proporcionando al menos un fragmento de ADN fuente bicatenario 202 y adaptadores 204 y 206. Los adaptadores 204 y 206 son similares a los adaptadores ilustrados en la Figura 2A(iii)-(iv). El adaptador 204 tiene una secuencia adaptadora P5 y un UMI α1 físico en su brazo 5'. El adaptador 204 también tiene una secuencia adaptadora P7' y un UMI α2 físico en su brazo 3'. El adaptador 206 tiene una secuencia adaptadora P5 y un UMI físico β2 en su brazo 5', y una secuencia adaptadora P7' y un UMI físico β1 en su brazo 3'. El proceso procede ligando el adaptador 204 y el adaptador 206 al fragmento 202, obteniéndose el producto de ligamiento 208. El proceso procede desnaturalizando el producto de ligamiento 208, dando lugar a un fragmento desnaturalizado monocatenario 212. Mientras tanto, una mezcla de reacción a menudo incluye adaptadores residuales en esta etapa. Dado que incluso si el proceso ya ha implicado la eliminación de adaptadores sobreabundantes, tales como, por ejemplo, utilizando bolas de Inmovilización Reversible en Fase Sólida (SPRI, solid-phase reversible immobilization), todavía quedan algunos adaptadores en la mezcla de reacción. Tal adaptador sobrante se ilustra como adaptador 210, que es similar al adaptador 206, excepto por que el adaptador 210 tiene UMI físicos γ1 y γ2 en sus brazos 3' y 7', respectivamente. La condición de desnaturalización que produce el fragmento desnaturalizado 212 también produce un oligonucleótido adaptador desnaturalizado 214, que tiene un UMI γ2 físico cerca de su secuencia adaptadora P5.

El fragmento adaptador monocatenario 214 se hibrida después con el fragmento de ADN catenario de señal 212, y un proceso de PCR amplía el fragmento adaptador monocatenario 214 para producir un inserto intermedio 216 que es complementario al fragmento de ADN 212. Durante los diversos ciclos de amplificación por PCR, los fragmentos adaptadores intermedios 218, 220 y 222 pueden ser el resultado de las extensiones de la PCR de las cadenas P7' de los adaptadores, incluidos los diferentes UMI físicos δ, ε y ζ. Los fragmentos adaptadores intermedios 218, 220 y 222 tienen todos la secuencia P7' en el extremo 5' y, respectivamente, tienen UMI físicos δ, ε y ζ. En los ciclos de PCR subsiguientes, los fragmentos adaptadores intermedios 218, 220 y 222 pueden hibridar con el fragmento intermedio 216 o sus amplicones, porque el extremo 3' de los fragmentos adaptadores intermedios 218, 220 y 222 es complementario a la región 217 del inserto intermedio 216. La extensión por PCR de los fragmentos hibridados produce fragmentos de ADN monocatenarios 224, 226 y 228. Los fragmentos de ADN 224, 226 y 228 están etiquetados con tres UMI físicos diferentes (δ, ε y ζ) en el extremo 5', y un UMI físico γ2 en el extremo 3', lo que indica un "salto de UMI", donde diferentes UMI están unidos a secuencias de nucleótidos derivadas del mismo fragmento de ADN 202.

En algunas implementaciones de la exposición, el uso de adaptadores que tienen UMI físicos en ambas cadenas de la región bicatenaria de los adaptadores, tales como los adaptadores de las Figuras 2A(v)-(vi), puede evitar o reducir el salto de UMI. Esto puede deberse al hecho de que los UMI físicos de un adaptador en la región bicatenaria son diferentes de los UMI físicos de todos los demás adaptadores. Esto ayuda a reducir la complementariedad entre los oligonucleótidos adaptadores intermedios y los fragmentos intermedios, evitando así la hibridación como la mostrada para el oligonucleótido intermedio 222 y el fragmento intermedio 220, reduciendo o previniendo así el salto de UMI.

Agrupamiento de lecturas y obtención de secuencias consenso

En diversas implementaciones que utilizan UMI, múltiples lecturas de secuencias que tienen los mismos UMI se agrupan para obtener una o más secuencias de consenso, que después se utilizan para determinar la secuencia de una molécula de ADN fuente. Pueden generarse múltiples lecturas distintas a partir de instancias distintas de la misma molécula de ADN fuente, y estas lecturas pueden compararse para producir una secuencia consenso como se describe en la presente memoria. Las instancias pueden generarse amplificando una molécula de ADN fuente antes de la secuenciación, de forma que se realicen distintas operaciones de secuenciación en distintos productos de amplificación, cada uno de los cuales comparte la secuencia de la molécula de ADN fuente. Por supuesto, la amplificación puede introducir errores, de modo que las secuencias de los distintos productos de amplificación

presenten diferencias. En el contexto de algunas tecnologías de secuenciación, tales como la secuenciación por síntesis de Illumina, una molécula de ADN fuente o un producto de amplificación de la misma forma un grupo de moléculas de ADN vinculadas a una región de una cubeta de lectura. Las moléculas del grupo proporcionan colectivamente una lectura. Normalmente, se necesitan al menos dos lecturas para obtener una secuencia consenso. Las profundidades de secuenciación de 100, 1000 y 10.000 son ejemplos de profundidades de secuenciación útiles en las realizaciones divulgadas para crear lecturas consenso para frecuencias alélicas bajas (p. ej., aproximadamente el 1 % o menos).

En algunas implementaciones, los nucleótidos que son coherentes en el 100 % de las lecturas que comparten un UMI o combinación de UMI se incluyen en la secuencia consenso. En otras implementaciones, el criterio de consenso puede ser inferior al 100 %. Por ejemplo, se puede utilizar un criterio de consenso del 90 %, lo que significa que los pares de bases que existen en el 90 % o más de las lecturas del grupo se incluyen en la secuencia consenso. En diversas implementaciones, el criterio de consenso puede establecerse en aproximadamente el 30 %, aproximadamente el 40 %, aproximadamente el 50 %, aproximadamente el 60 %, aproximadamente el 70 %, aproximadamente el 80 %, aproximadamente el 90 %, aproximadamente el 95 % o aproximadamente el 100 %.

Agrupamiento mediante UMI físicos y UMI virtuales

Se pueden utilizar varias técnicas para agrupar lecturas que incluyan varios UMI. En algunas implementaciones, las lecturas que comparten un UMI físico común pueden agruparse para obtener una secuencia consenso. En algunas implementaciones, si el UMI físico común es un UMI aleatorio, el UMI aleatorio puede ser lo suficientemente único como para identificar una molécula fuente particular de un fragmento de ADN en una muestra. En otras implementaciones, si el UMI físico común es un UMI no aleatorio, el UMI puede no ser lo suficientemente único por sí mismo para identificar una molécula fuente concreta. En cualquier caso, un UMI físico puede combinarse con un UMI virtual para proporcionar un índice de la molécula de origen.

En el ejemplo de flujo de trabajo descrito anteriormente y representado en las Figuras 1B, 3A y 4, algunas lecturas incluyen UMI α-ρ-ϕ, mientras que otras incluyen UMI β-ϕ-ρ. El UMI físico α produce lecturas que tienen α. Si todos los adaptadores utilizados en un flujo de trabajo tienen diferentes UMI físicos (p. ej., diferentes UMI aleatorios), es probable que todas las lecturas que tengan α en la región del adaptador procedan de la misma cadena del fragmento de ADN. Del mismo modo, el UMI físico β produce lecturas que tienen β, todas ellas derivadas de la misma cadena complementaria del fragmento de ADN. Por lo tanto, es útil agrupar todas las lecturas que incluyan α para obtener una secuencia consenso, y agrupar todas las lecturas que incluyan β para obtener otra secuencia consenso. Esto se ilustra como el agrupamiento de primer nivel en las Figuras 4B-4C. Dado que todas las lecturas de un grupo proceden del mismo polinucleótido fuente de una muestra, los pares de bases incluidos en la secuencia consenso probablemente reflejen la verdadera secuencia del polinucleótido fuente, mientras que un par de bases excluido de la secuencia consenso probablemente refleje una variación o error introducido en el flujo de trabajo.

Además, los UMI virtuales ρ y ϕ pueden proporcionar información para determinar que las lecturas que incluyen uno o ambos UMI virtuales proceden del mismo fragmento de ADN fuente. Dado que los UMI virtuales ρ y ϕ son internos a los fragmentos de ADN fuente, en la práctica, la explotación de los UMI virtuales no añade sobrecarga a la preparación o secuenciación. Tras obtener las secuencias de los UMI físicos a partir de las lecturas, pueden determinarse una o más subsecuencias de las lecturas como UMI virtuales. Si los UMI virtuales incluyen suficientes pares de bases y tienen la misma ubicación relativa en las lecturas, pueden identificar de forma única las lecturas como derivadas del fragmento de ADN fuente. Por lo tanto, las lecturas que tengan uno o ambos UMI virtuales ρ y ϕ pueden agruparse para obtener una secuencia consenso. La combinación de UMI virtuales y UMI físicos puede proporcionar información para guiar un agrupamiento de segundo nivel cuando solo se asigna un UMI físico a una secuencia de consenso de primer nivel de cada cadena, según muestra la Figura 3A y las Figuras 4A-4C. Sin embargo, en algunas implementaciones, este segundo nivel de agrupamiento utilizando UMI virtuales puede ser difícil si hay moléculas de ADN de entrada sobreabundantes o la fragmentación no es aleatoria.

En realizaciones alternativas, las lecturas que tienen dos UMI físicos en ambos extremos, tales como las que se muestran en la Figura 3B y en las Figuras 4D y 4E, pueden agruparse en un agrupamiento de segundo nivel basado en una combinación de los UMI físicos y los UMI virtuales. Esto resulta especialmente útil cuando los UMI físicos son demasiado cortos para identificar de forma inequívoca los fragmentos de ADN fuente sin utilizar los UMI virtuales. En estas realizaciones, puede implementarse el agrupamiento de segundo nivel, con UMI físicos dúplex según muestra la Figura 3B, mediante el agrupamiento de lecturas consenso α-ρ-ϕ-β y lecturas consenso β-ϕ-ρ-α de la misma molécula de ADN, obteniendo de esta manera una secuencia de consenso que incluye nucleótidos coherentes entre todas las lecturas.

Utilizando UMI y el esquema de agrupamiento descrito en la presente memoria, diversas realizaciones pueden suprimir diferentes fuentes de error que afectan a la secuencia determinada de un fragmento, incluso si el fragmento incluye alelos con frecuencias alélicas muy bajas. Se agrupan las lecturas que comparten los mismos UMI (físicos y/o virtuales). Al agrupar las lecturas agrupadas, se pueden eliminar las variantes (SNV y pequeñas indel) debidas a errores de PCR, preparación de genotecas, agrupación y secuenciación. Las Figuras 4A-4E ilustran cómo un método como el divulgado en un ejemplo de flujo de trabajo puede suprimir diferentes fuentes de error en la determinación de

la secuencia de un fragmento de ADN bicatenario. Las lecturas ilustradas incluyen UMI α-ρ-ϕ o β-ϕ-ρ en las Figuras 3A y 4A-4C, y UMI α-ρ-ϕ-β o β-ϕ-ρ-α en las Figuras 3B, 4D y 4E. Los UMI α y β son UMI físicos de una unidad de plexado en las Figuras 3A y 4A-4C. Los UMI α y β son UMI dúplex en las Figuras 3B, 4D y 4E. Los UMI virtuales ρ y ϕ están situados en los extremos de un fragmento de ADN.

El método que utiliza UMI físicos de una unidad de plexado, como se muestra en las Figuras 4A-4C, implica primero el agrupamiento de lecturas que tienen el mismo UMI físico α o β, ilustrado como agrupamiento de primer nivel. El agrupamiento de primer nivel obtiene una secuencia consenso α para las lecturas que tienen el UMI físico α, que proceden de una cadena del fragmento bicatenario. El agrupamiento de primer nivel también obtiene una secuencia consenso β para las lecturas que tienen el UMI físico β, que proceden de otra cadena del fragmento bicatenario. En un agrupamiento de segundo nivel, el método obtiene una tercera secuencia consenso a partir de la secuencia consenso α y la secuencia consenso β. La tercera secuencia consenso refleja pares de bases consenso de lecturas que tienen los mismos UMI virtuales dúplex ρ y ϕ, que proceden de dos cadenas complementarias del fragmento fuente. Por último, la secuencia del fragmento de ADN bicatenario se determina como la tercera secuencia consenso.

El método que utiliza UMI físicos dúplex como se muestra en las Figuras 4D-4E implica primero el agrupamiento de lecturas que tienen los UMI físicos α y β con un orden α→β en la dirección 5'-3', ilustrado como agrupamiento de primer nivel. La agrupación de primer nivel obtiene una secuencia de consenso α→β para lecturas que tienen los UMI físicos α y β, de los cuales se derivan las lecturas de una primera cadena del fragmento bicatenario. El colapso de primer nivel también obtiene una secuencia consenso β-α para las lecturas que tienen los UMI físicos β y α con un orden β→α en la dirección 5'-3', que proceden de una segunda cadena complementaria a la primera cadena del fragmento bicatenario. En un segundo nivel de agrupamiento, el método obtiene una tercera secuencia consenso a partir de la secuencia consenso α-β y la secuencia consenso β-α. La tercera secuencia consenso refleja pares de bases consenso de lecturas que tienen los mismos UMI virtuales dúplex ρ y ϕ, que proceden de dos cadenas del fragmento. Por último, la secuencia del fragmento de ADN bicatenario se determina como la tercera secuencia consenso.

La Figura 4A ilustra cómo un agrupamiento de primer nivel puede suprimir los errores de secuenciación. Los errores de secuenciación se producen en la plataforma de secuenciación después de la preparación de la muestra y la biblioteca (p. ej., amplificación por PCR). Los errores de secuenciación pueden introducir diferentes bases erróneas en diferentes lecturas. Las bases verdaderas positivas se ilustran con letras sólidas, mientras que las bases falsas positivas se ilustran con letras sombreadas. Los nucleótidos falsos positivos en diferentes lecturas de la familia α-ρ-ϕ se han excluido de la secuencia consenso α. El nucleótido verdadero positivo "A" ilustrado en los extremos izquierdos de las lecturas de la familia α-ρ-ϕ se mantiene para la secuencia de consenso α. Similarmente, los nucleótidos falsos positivos en diferentes lecturas de la familia β-ϕ-ρ se han excluido de la secuencia de consenso β, conservando el nucleótido verdadero positivo "A". Como se ilustra aquí, el agrupamiento de primer nivel puede eliminar eficazmente los errores de secuenciación. La Figura 4A también muestra un agrupamiento opcional de segundo nivel basado en los UMI virtuales ρ y ϕ. Este agrupamiento de segundo nivel puede suprimir aún más los errores, como se ha explicado anteriormente, pero tales errores no se ilustran en la Figura 4A.

Los errores de PCR se producen antes de la amplificación del agrupamiento. Por lo tanto, un par de bases erróneo introducido en un ADN monocatenario por el proceso de PCR puede amplificarse durante la amplificación del agrupamiento, apareciendo así en múltiples grupos y lecturas. Como se ilustra en la Figura 4B y en la Figura 4D, en muchas lecturas puede aparecer un par de bases falso positivo introducido por un error de PCR. La base "T" en las lecturas de la familia α-ρ-ϕ (Figura 4B) o α-β (Figura 4D) y la base "C" en las lecturas de la familia β-ϕ-ρ (Figura 4B) o β-α (Figura 4D) son tales errores de PCR. En cambio, los errores de secuenciación mostrados en la Figura 4A aparecen en una o unas pocas lecturas de la misma familia. Dado que los errores de secuenciación mediante la PCR aparecen en muchas lecturas de la familia, un agrupamiento de primer nivel de las lecturas en una cadena no elimina los errores de la PCR, incluso si el primer agrupamiento elimina los errores de secuenciación (p. ej. G y A eliminados de la familia α-ρ-ϕ en la Figura 4B y de la familia α-β en la Figura 4D). Sin embargo, dado que un error de PCR se introduce en un ADN monocatenario, la cadena complementaria del fragmento de origen y las lecturas procedentes del mismo no suelen tener el mismo error de PCR. Por lo tanto, el agrupamiento de segundo nivel basado en las lecturas de las dos cadenas del fragmento de origen puede eliminar eficazmente los errores de PCR, como se muestra en la parte inferior de las Figuras 4B y 4D.

En algunas plataformas de secuenciación, los errores de homopolímero se producen para introducir pequeños errores de indel en homopolímeros de nucleótidos únicos repetidos. Las Figuras 4C y 4E ilustran la corrección de errores de homopolímeros utilizando los métodos descritos en la presente memoria. En las lecturas de las familias α-ρ-ϕ (Figura 4C) o α-ρ-ϕ-β (Figura 4E), se han eliminado dos nucleótidos "T" de la segunda lectura desde arriba y un nucleótido "T" de la tercera lectura desde arriba. En las lecturas de la familia β-ϕ-ρ (Figura 4C) o β-ϕ-ρ-α (Figura 4E), se ha insertado un nucleótido "T" en la primera lectura desde la parte superior. De forma similar al error de secuenciación ilustrado en la Figura 4A, los errores de homopolímero se producen tras la amplificación PCR, por lo que diferentes lecturas tienen diferentes errores de homopolímero. Como resultado, el agrupamiento de primer nivel puede eliminar eficazmente los errores de indel.

Las secuencias consenso pueden obtenerse agrupando lecturas que tengan uno o más UMI comunes no aleatorios y uno o más UMI virtuales comunes. Además, la información de posición también puede utilizarse para obtener secuencias consenso, como se describe a continuación.

Agrupamiento por posición

En algunas implementaciones, las lecturas se procesan para alinearse con una secuencia de referencia para determinar las ubicaciones de alineación de las lecturas en la secuencia de referencia (localización). Sin embargo, en algunas implementaciones no ilustradas anteriormente, la localización se consigue mediante el análisis de similitud k-mero y la alineación lectura-lectura. Esta segunda implementación tiene dos ventajas: en primer lugar, puede agrupar (corregir errores de) las lecturas que no coinciden con la referencia, debido a diferencias haplotípicas o translocaciones, y en segundo lugar, no depende de un algoritmo de alineación, eliminando de esta manera la posibilidad de artefactos inducidos por el alineador (errores en el alineador). En algunas implementaciones, las lecturas que comparten la misma información de localización pueden agruparse para obtener secuencias consenso para determinar la secuencia de los fragmentos de ADN fuente. En algunos contextos, el proceso de alineamiento también se denomina proceso de cartografiado. Las lecturas de secuencia se someten a un proceso de alineamiento para cartografiarlas a una secuencia de referencia. Pueden usarse diversas herramientas y algoritmos de alineación para alinear las lecturas a la secuencia de referencia como se describe en cualquier otro sitio de la exposición. Como es habitual, en los algoritmos de alineamiento, algunas lecturas se alinean con éxito con la secuencia de referencia, mientras que otras pueden no alinearse con éxito o pueden quedar mal alineadas con la secuencia de referencia. Las lecturas que se alinean sucesivamente con la secuencia de referencia se asocian con sitios en la secuencia de referencia. Las lecturas alineadas y sus sitios asociados también se denominan marcadores de secuencia. Algunas lecturas de secuencia que contienen un gran número de repeticiones tienden a ser más difíciles de alinear con la secuencia de referencia. Cuando una lectura se alinea con una secuencia de referencia con una serie de bases no coincidentes por encima de un determinado criterio, la lectura se considera mal alineada. En diversas realizaciones, las lecturas se consideran mal alineadas cuando están alineadas con al menos aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, o 10 emparejamientos erróneos. En otras realizaciones, las lecturas se consideran mal alineadas cuando están alineadas con al menos aproximadamente un 5 % de emparejamientos erróneos. En otras realizaciones, las lecturas se consideran mal alineadas cuando están alineadas con al menos aproximadamente un 10 %, 15 % o 20 % de bases no coincidentes.

En algunas implementaciones, los métodos divulgados combinan información de posición con información de UMI físicos para indexar moléculas fuente de fragmentos de ADN. Las lecturas de secuencias que compartan una misma posición de lectura y un mismo UMI físico aleatorio o no aleatorio pueden agruparse para obtener una secuencia de consenso para determinar la secuencia de un fragmento o porción del mismo. En algunas implementaciones, las lecturas de secuencias que comparten la misma posición de lectura, el mismo UMI físico no aleatorio y un UMI físico aleatorio pueden fusionarse para obtener una secuencia consenso. En tales implementaciones, el adaptador puede incluir tanto un UMI físico no aleatorio como un UMI físico aleatorio. En algunas implementaciones, las lecturas de secuencias que comparten la misma posición de lectura y el mismo UMI virtual pueden agruparse para obtener una secuencia consenso.

La información sobre la posición de lectura puede obtenerse mediante diferentes técnicas. Por ejemplo, en algunas implementaciones, las coordenadas genómicas pueden utilizarse para proporcionar información sobre la posición de la lectura. En algunas implementaciones, la posición en una secuencia de referencia con la que se alinea una lectura puede utilizarse para proporcionar información sobre la posición de la lectura. Por ejemplo, las posiciones de inicio y parada de una lectura en un cromosoma pueden utilizarse para proporcionar información sobre la posición de la lectura. En algunas implementaciones, las posiciones de lectura se consideran iguales si tienen idéntica información de posición. En algunas implementaciones, las posiciones de lectura se consideran iguales si la diferencia entre la información de posición es menor que un criterio definido. Por ejemplo, dos lecturas con posiciones genómicas de inicio que difieran en menos de 2, 3, 4 o 5 pares de bases pueden considerarse lecturas con la misma posición de lectura. En otras implementaciones, las posiciones de lectura se consideran iguales si su información de posición se puede convertir y hacer coincidir en un espacio de posición concreto. Puede proporcionarse una secuencia de referencia antes de la secuenciación (por ejemplo, puede ser una secuencia genómica humana conocida y ampliamente utilizada) o puede determinarse a partir de las lecturas obtenidas durante la secuenciación de la muestra.

Independientemente de la plataforma y el protocolo de secuenciación específicos, al menos una porción de los ácidos nucleicos contenidos en la muestra se secuencia para generar decenas de miles, cientos de miles, o millones de lecturas de secuencia, p. ej., lecturas de 100 pb. En algunas realizaciones, las lecturas de secuencia comprenden aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 36 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130 pb, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb,

aproximadamente 450 pb, aproximadamente 500 pb, aproximadamente 800 pb, aproximadamente 1000 pb o aproximadamente 2000 pb.

En algunas realizaciones, las lecturas se alinean con un genoma de referencia, p. ej., hg19. En otras realizaciones, las lecturas se alinean con una porción de un genoma de referencia, p. ej., un cromosoma o un segmento cromosómico. Las lecturas que se asignan de manera única al genoma de referencia se conocen como marcadores de secuencia. En una realización, se obtienen al menos aproximadamente 3 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 5 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 8 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 10 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 15 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 20 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 30 x 10^6 etiquetas de secuencia calificadas, al menos aproximadamente 40 x 10^6 etiquetas de secuencia calificadas o al menos aproximadamente 50 x 10^6 etiquetas de secuencia calificadas a partir de lecturas que se mapean de forma exclusiva a un genoma de referencia.

Aplicaciones

En diversas aplicaciones, las estrategias de corrección de errores tal como se divulgan en la presente memoria pueden proporcionar uno o más de los siguientes beneficios: (i) detectar mutaciones somáticas de muy baja frecuencia alélica, (ii) reducir el tiempo de ciclo mitigando los errores de fase de hebra retrasada/fase de hebra adelantada y/o (iii) aumentar la longitud de lectura potenciando la calidad de las llamadas de bases en la parte posterior de las lecturas, etc. Las aplicaciones y justificaciones de la detección de las mutaciones somáticas de baja frecuencia alélica se han analizado anteriormente.

En ciertas realizaciones, las técnicas descritas en la presente memoria pueden permitir la llamada fiable de alelos con frecuencias de aproximadamente el 2 % o menos, o aproximadamente el 1 % o menos, o aproximadamente el 0,5 % o menos. Tales bajas frecuencias son comunes en el ADNlc procedente de células tumorales en un paciente oncológico. En algunas realizaciones, las técnicas descritas en la presente memoria pueden permitir la identificación de cepas raras en muestras metagenómicas, así como la detección de variantes raras en poblaciones de virus o de otro tipo cuando, por ejemplo, un paciente ha sido infectado por múltiples cepas de virus y/o ha sido sometido a tratamiento médico.

En ciertas realizaciones, las técnicas descritas en la presente memoria pueden permitir un tiempo de ciclo de química de secuenciación más corto. La reducción del tiempo de ciclo aumenta los errores de secuenciación, que pueden corregirse mediante el método descrito anteriormente.

En algunas implementaciones que implican UMI, las lecturas largas pueden obtenerse mediante secuenciación de ambos extremos utilizando longitudes de lectura asimétricas para un par de lecturas de ambos extremos (PE, por sus siglas en inglés) a partir de los dos extremos de un segmento. Por ejemplo, un par de lecturas que tengan 50 pb en una lectura de ambos extremos y 500 pb en otra lectura de ambos extremos pueden "unirse" con otro par de lecturas para producir una lectura larga de 1000 pb. Estas implementaciones pueden proporcionar una mayor velocidad de secuenciación para determinar fragmentos largos de bajas frecuencias alélicas.

La Figura 5 ilustra esquemáticamente un ejemplo para obtener eficientemente lecturas largas de extremos emparejados en este tipo de aplicaciones aplicando UMI físicos y UMI virtuales. Las bibliotecas de ambas cadenas de los mismos fragmentos de ADN se agrupan en la celda de flujo. El tamaño del inserto de la librería es superior a 1 Kb. La secuenciación se realiza con longitudes de lectura asimétricas (p. ej., Lectura1 = 500 pb, Lectura2 = 50 pb), para garantizar la calidad de las lecturas largas de 500 pb. Al unir dos cadenas, se pueden crear lecturas PE de 1000 pb de longitud con solo 500 + 50 pb de secuenciación.

Muestras

Las muestras que se usan para determinar la secuencia de fragmento de ADN pueden incluir muestras tomadas de cualquier célula, líquido, tejido u órgano, incluidos ácidos nucleicos en los que se van a determinar secuencias de interés. En algunas realizaciones que implican el diagnóstico de cánceres, el ADN tumoral circulante puede obtenerse de un líquido corporal del sujeto, p. ej., sangre o plasma. En algunas realizaciones que implican diagnósticos en fetos, es ventajoso obtener ácidos nucleicos libres circulantes, p. ej., ADN libre circulante (ADNlc), a partir de fluidos corporales maternos. Los ácidos nucleicos libres circulantes, incluido el ADN libre circulante, pueden obtenerse mediante diversos métodos conocidos en la técnica a partir de muestras biológicas que incluyen, aunque no de forma limitativa, plasma, suero y orina (véase, p. ej., Fan y col., Proc Natl Acad Sci 105:16266-16271 [2008]; Koide y col., Prenatal Diagnosis 25:604-607 [2005]; Chen y col., Nature Med. 2: 1033-1035 [1996]; Lo y col., Lancet 350: 485-487 [1997]; Botezatu y col., Clin Chem. 46: 1078-1084, 2000; y Su y col., J Mol. Diagn. 6: 101-107 [2004]).

En diversas realizaciones, los ácidos nucleicos (por ejemplo, ADN o ARN) presentes en la muestra pueden enriquecerse específicamente o no específicamente antes de su uso (por ejemplo, antes de preparar una biblioteca de secuenciación). El enriquecimiento no específico de ADN de muestra se refiere a la amplificación del genoma completo de los fragmentos de ADN genómico de la muestra que pueden usarse para aumentar el nivel del ADN de

la muestra antes de preparar una biblioteca de secuenciación de ADNlc. Los métodos para la amplificación del genoma completo son conocidos en la técnica. La PCR cebada con oligonucleótidos degenerados (DOP), la técnica de PCR de extensión por cebadores (PEP) y la amplificación por desplazamiento múltiple (MDA) son ejemplos de métodos de amplificación del genoma completo. En algunas realizaciones, no se ha enriquecido el ADN de la muestra.

La muestra que incluye los ácidos nucleicos a los que se aplican los métodos descritos en la presente memoria típicamente incluye una muestra biológica ("muestra de ensayo"), como se ha descrito anteriormente. En algunas realizaciones, los ácidos nucleicos que se van a secuenciar se purifican o aíslan mediante cualquiera de diversos métodos bien conocidos.

Por consiguiente, en determinadas realizaciones la muestra incluye o consiste esencialmente en un polinucleótido purificado o aislado, o puede incluir muestras tales como una muestra de tejido, una muestra de líquido biológico, una muestra celular y lo similar. Las muestras de líquido biológico adecuadas incluyen, aunque no de forma limitativa, plasma, suero, sudor, lágrimas, esputo, orina, esputo, exudado del oído, linfa, saliva, líquido cefalorraquídeo, desechos, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, ascitis, leche, secreciones de las vías respiratorias, el tubo digestivo y las vías genitourinarias, líquido amniótico, leche y muestras de leucoforesis. En algunas realizaciones, la muestra es una muestra que se puede obtener fácilmente mediante procedimientos no invasivos, p. ej., sangre, plasma, suero, sudor, lágrimas, esputo, orina, heces, esputo, secreción del oído, saliva o heces. En determinadas realizaciones, la muestra es una muestra de sangre periférica o las fracciones de plasma y/o de suero de una muestra de sangre periférica. En otras realizaciones, la muestra biológica es un hisopo o frotis, una muestra de biopsia o un cultivo celular. En otra realización, la muestra es una mezcla de dos o más muestras biológicas, p. ej., una muestra biológica puede incluir dos o más de una muestra de fluido biológico, una muestra de tejido y una muestra de cultivo celular. Como se utilizan en la presente memoria, los términos "sangre", "plasma" y "suero" abarcan expresamente fracciones o porciones procesadas de los mismos. Similarmente, cuando una muestra se toma de una biopsia, un hisopo, un frotis, etc., la "muestra" abarca expresamente una fracción o porción procesada obtenida de la biopsia, el hisopo, el frotis, etc.

En determinadas realizaciones, las muestras pueden obtenerse de fuentes, que incluyen, aunque no de forma limitativa, muestras de diferentes individuos, muestras de diferentes etapas de desarrollo del mismo o diferentes individuos, muestras de diferentes individuos enfermos (p. ej., individuos que se sospecha tienen un trastorno genético), individuos normales, muestras obtenidas en diferentes etapas de una enfermedad en un individuo, muestras obtenidas de un individuo sometido a diferentes tratamientos para una enfermedad, muestras de individuos sometidos a diferentes factores ambientales, muestras de individuos con predisposición a una patología, muestras individuales con exposición a un agente de enfermedad infecciosa, y lo similar.

En una realización ilustrativa, pero no limitativa, la muestra es una muestra materna que se obtiene de una hembra embarazada, por ejemplo, una mujer embarazada. En este caso, la muestra puede analizarse usando los métodos descritos en la presente memoria para proporcionar un diagnóstico prenatal de posibles anomalías cromosómicas en el feto. La muestra materna puede ser una muestra de tejido, una muestra de fluido biológico o una muestra celular. Un líquido biológico incluye, como ejemplos no limitativos, sangre, plasma, suero, sudor, lágrimas, esputo, orina, esputo, exudado del oído, linfa, saliva, líquido cefalorraquídeo, desechos, suspensión de médula ósea, flujo vaginal, lavado transcervical, líquido cerebral, ascitis, leche, secreciones de las vías respiratorias, el tubo digestivo y las vías genitourinarias, y muestras de leucoforesis.

En determinadas realizaciones, también se pueden obtener muestras de tejidos, células u otras fuentes que contienen polinucleótidos cultivados *in vitro*. Las muestras cultivadas pueden tomarse de fuentes que incluyen, aunque no de forma limitativa, cultivos (p. ej., tejido o células) mantenidas en diferentes medios y condiciones (p. ej., pH, presión o temperatura), cultivos (p. ej., tejido o células) mantenidos durante diferentes períodos de longitud, cultivos (p. ej., tejido o células) tratados con diferentes factores o reactivos (p. ej., un candidato a fármaco, o un modulador), o cultivos de diferentes tipos de tejido y/o células.

Los métodos para aislar ácidos nucleicos de fuentes biológicas son bien conocidos y diferirán dependiendo de la naturaleza de la fuente. Un experto en la técnica puede aislar fácilmente ácidos nucleicos de una fuente según sea necesario para el método descrito en la presente memoria. En algunos casos, puede ser ventajoso fragmentar las moléculas de ácido nucleico en la muestra de ácido nucleico. La fragmentación puede ser al azar, o puede ser específica, como se logra, por ejemplo, usando digestión con endonucleasas de restricción. Los métodos de fragmentación al azar son bien conocidos en la técnica e incluyen, por ejemplo, digestión con ADNasa limitada, tratamiento con álcali y cizallamiento físico.

Preparación de la biblioteca de secuenciación

En diversas realizaciones, la secuenciación puede realizarse en diversas plataformas de secuenciación que requieren la preparación de una biblioteca de secuenciación. La preparación implica normalmente fragmentar el ADN (tratamiento con ultrasonidos, nebulización o cizalladura), seguido de la reparación del ADN y el refinado de los extremos (extremos romos o saliente de A) y la ligadura de adaptadores específicos para la plataforma. En una realización, los métodos descritos en la presente memoria pueden utilizar tecnologías de secuenciación de próxima

generación (NGS), que permiten secuenciar múltiples muestras individualmente como moléculas genómicas (es decir, secuenciación singleplex) o como muestras agrupadas que comprenden moléculas genómicas indexadas (por ejemplo, secuenciación multiplexada) en una sola tanda de secuenciación. Estos métodos pueden generar hasta varios miles de millones de lecturas de secuencias de ADN. En diversas realizaciones, las secuencias de ácidos nucleicos genómicos y/o de ácidos nucleicos genómicos indexados se pueden determinar usando, por ejemplo, las tecnologías de secuenciación de nueva generación (NGS) descritas en la presente memoria. En diversas realizaciones, el análisis de la cantidad masiva de datos de secuencia obtenidos usando NGS puede realizarse usando uno o más procesadores como se describe en la presente memoria.

En diversas realizaciones, el uso de tales tecnologías de secuenciación no implica la preparación de bibliotecas de secuenciación.

Sin embargo, en determinadas realizaciones, los métodos de secuenciación contemplados en la presente memoria implican la preparación de bibliotecas de secuenciación. En un planteamiento ilustrativo, la preparación de la biblioteca de secuenciación implica la producción de una colección al azar de fragmentos de ADN modificados con adaptador (p. ej., polinucleótidos) que están listos para ser secuenciados. Las bibliotecas de secuenciación de polinucleótidos pueden prepararse a partir de ADN o ARN, incluidos equivalentes, análogos de ADN o ADNc, por ejemplo, ADN o ADNc, que es ADN complementario o copia producido a partir de una cadena molde de ARN, por la acción de la transcriptasa inversa. Los polinucleótidos pueden originarse en forma bicatenaria (p. ej., ADNbc tales como fragmentos de ADN genómico, ADNc, productos de amplificación por PCR y similares) o, en determinadas realizaciones, los polinucleótidos pueden originarse en forma monocatenaria (p. ej., ADNmc, ARN, etc.) y se han convertido en forma de ADNbc. A modo de ilustración, en determinadas realizaciones, las moléculas de ARNm monocatenario pueden copiarse en ADNc bicatenarios adecuados para su uso en la preparación de una biblioteca de secuenciación. La secuencia precisa de las moléculas de polinucleótidos primarios generalmente no es parte esencial del método de preparación de bibliotecas, y puede ser conocida o desconocida. En una realización, las moléculas de polinucleótido son moléculas de ADN. Más particularmente, en determinadas realizaciones, las moléculas de polinucleótido representan todo el complemento genético de un organismo o sustancialmente el complemento genético completo de un organismo, y son moléculas de ADN genómico (p. ej., ADN celular, ADN libre de células (ADNlc), etc.), que de forma típica incluyen tanto secuencia de intrones como secuencia de exones (secuencia codificante), así como secuencias reguladoras no codificantes tales como secuencias promotoras y potenciadoras. En ciertas realizaciones, las moléculas de polinucleótidos primarios comprenden moléculas de ADN genómico humano, p. ej., moléculas de ADNlc presentes en la sangre periférica de una sujeto embarazada.

La preparación de bibliotecas de secuenciación para algunas plataformas de secuenciación de NGS se facilita mediante el uso de polinucleótidos que comprenden un intervalo específico de tamaños de fragmentos. La preparación de tales bibliotecas implica típicamente la fragmentación de polinucleótidos grandes (p. ej., ADN genómico celular) para obtener polinucleótidos en el intervalo de tamaño deseado.

En los métodos y sistemas de secuenciación expuestos en la presente memoria pueden utilizarse lecturas de ambos extremos. La longitud del fragmento o inserto es mayor que la longitud de lectura y en ocasiones, mayor que la suma de las longitudes de las dos lecturas.

En algunas realizaciones ilustrativas, se obtienen el ácido o los ácidos nucleicos de muestra como ADN genómico, que se somete a fragmentación en fragmentos más largos de aproximadamente 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000 o 5000 pares de bases, a los que pueden aplicarse fácilmente los métodos de NGS. En algunas realizaciones, las lecturas de ambos extremos se obtienen de insertos de aproximadamente 100–5000 pb. En algunas realizaciones, los insertos tienen una longitud de aproximadamente 100-1000 pb. Estos se implementan a veces como lecturas de ambos extremos de inserto corto normales. En algunas realizaciones, los insertos tienen una longitud de aproximadamente 1000–5000 pb. Estos a veces se implementan como lecturas emparejadas por parejas de inserto largo como se ha descrito anteriormente.

En algunas implementaciones, se diseñan insertos largos para evaluar secuencias muy largas. En algunas implementaciones, las lecturas de pares de parejas pueden aplicarse para obtener lecturas que están separadas por miles de pares de bases. En estas implementaciones, los insertos de los fragmentos varían de cientos a miles de pares de bases, con dos adaptadores de unión a biotina en los dos extremos de un inserto. A continuación, los adaptadores de unión a biotina unen los dos extremos del inserto, formándose una molécula circularizada, que a continuación se fragmenta adicionalmente. Para la secuenciación en una plataforma diseñada para secuenciar fragmentos más cortos se selecciona un subfragmento que incluye los adaptadores de unión a biotina y los dos extremos del inserto original.

La fragmentación se puede lograr mediante cualquiera de entre diversos métodos conocidos por los expertos en la técnica. Por ejemplo, la fragmentación puede lograrse mediante medios mecánicos incluidos, aunque no de forma limitativa, nebulización, sonicación e hidrocizallamiento. Sin embargo, la fragmentación mecánica típicamente escinde la cadena principal de ADN en enlaces C-O, P-O y C-C, que resulta en una mezcla heterogénea de extremos romos y salientes 3' y 5' con enlaces C-O, P-O y/ C-C romos (véase, p. ej., Alnemri y Liwack, J Biol. Chem 265:17323-17333 [1990]; Richards y Boyer, J Mol Biol 11:327–240 [1965]) que puede ser necesario reparar, ya que pueden carecer del

fosfato 5' necesario para las posteriores reacciones enzimáticas, p. ej., ligamiento de adaptadores de secuenciación, que son necesarios para preparar el ADN para la secuenciación.

Por el contrario, el ADNlc existe de forma típica como fragmentos de menos de aproximadamente 300 pares de bases y, por consiguiente, no es de forma típica necesaria la fragmentación para generar una biblioteca de secuenciación usando muestras de ADNlc.

De forma típica, si los polinucleótidos se fragmentan a la fuerza (p. ej., se fragmentan *in vitro*), o existen naturalmente como fragmentos, se convierten en ADN de extremos romos que tiene fosfatos en 5' e hidroxilo en 3'. Los protocolos estándar, p. ej., los protocolos de secuenciación que emplean, por ejemplo, la plataforma de Illumina como se describe en el ejemplo de flujo de trabajo anterior en referencia a las Figuras 1A y 1B, indican al usuario que deben realizar la reparación de los extremos de la muestra de ADN, la purificación de los productos con los extremos reparados antes de la adenilación o la adición de la cola de dA en los extremos 3' y la purificación de los productos tras la adición de la cola de dA antes de los pasos de ligamiento de adaptadores de la preparación de bibliotecas.

Diversas realizaciones de métodos de preparación de bibliotecas de secuencias descritas en la presente memoria no suponen la necesidad de realizar una o más de las etapas típicamente exigidas por los protocolos estándar para obtener un producto de ADN modificado que puede secuenciarse mediante NGS. Un método abreviado (método ABB), un método de una etapa y un método de dos etapas son ejemplos de métodos para la preparación de una biblioteca de secuenciación, que se pueden encontrar en la solicitud de patente 13/555,037 presentada el 20 de julio de 2012, que se menciona en la presente memoria como referencia en su totalidad.

Métodos de secuenciación

Los métodos y aparatos descritos en la presente memoria pueden emplear tecnología de secuenciación de nueva generación (NGS), que permite la secuenciación masivamente en paralelo. En ciertas realizaciones, las plantillas de ADN amplificadas clonalmente o las moléculas de ADN individuales se secuencian de forma masivamente paralela dentro de una cubeta de lectura (p. ej., como se describe en Volkerding y col. Clin Chem 55:641-658 [2009]; Metzker M Nature Rev 11:31-46 [2010]). Las tecnologías de secuenciación de NGS incluyen, aunque no de forma limitativa, pirosecuenciación, secuenciación por síntesis con terminadores de colorante reversibles, secuenciación por ligadura de sonda de oligonucleótido y secuenciación de semiconductores de iones. El ADN de muestras individuales puede secuenciarse individualmente (es decir, secuenciación de una unidad de plexado) o el ADN de múltiples muestras puede agruparse y secuenciarse como moléculas genómicas indexadas (es decir, secuenciación multiplexada) en un solo experimento de secuenciación, para generar hasta varios cientos de millones de lecturas de secuencias de ADN. En la presente memoria se describen adicionalmente algunos ejemplos de tecnologías de secuenciación que pueden utilizarse para obtener la información de secuencia según el presente método.

En el comercio se dispone de algunas tecnologías de secuenciación, tales como la plataforma de secuenciación por hibridación de Affymetrix Inc. (Sunnyvale, CA) y las plataformas de secuenciación por síntesis de 454 Life Sciences (Bradford, CT), Illumina/Solexa (Hayward, CA) y Helicos Biosciences (Cambridge, MA) y la plataforma de secuenciación por ligamiento de Applied Biosystems (Foster City, CA), como se describe a continuación. Además de la secuenciación de moléculas individuales realizada mediante la secuenciación por síntesis de Helicos Biosciences, otras tecnologías de secuenciación de moléculas individuales incluyen, pero sin limitación, la tecnología SMRT™ de Pacific Biosciences, la tecnología ION TORRENT™, y la secuenciación de nanoporos desarrollada, por ejemplo, por Oxford Nanopore Technologies.

Aunque el método automatizado de Sanger se considera una tecnología de "primera generación", la secuenciación de Sanger, incluida la secuenciación automatizada de Sanger, también puede emplearse en los métodos descritos en la presente memoria. Los métodos de secuenciación adecuados adicionales incluyen, aunque no de forma limitativa, tecnologías de obtención de imágenes de ácidos nucleicos, p. ej., microscopía de fuerza atómica (AFM) o microscopía electrónica de transmisión (MET). Las tecnologías de secuenciación ilustrativas se describen con mayor detalle a continuación.

En algunas realizaciones, los métodos expuestos implican la obtención de información de secuencia para los ácidos nucleicos de la muestra de prueba mediante secuenciación masivamente paralela de millones de fragmentos de ADN utilizando la secuenciación por síntesis de Illumina y la química de secuenciación reversible basada en terminadores (p. ej., como se describe en Bentley y col., Nature 6:53-59 [2009]). La cadena molde de ADN puede ser ADN genómico, p. ej., ADN celular o ADNlc. En algunas realizaciones, el ADN genómico de células aisladas se usa como cadena molde, y se fragmenta en longitudes de varios cientos de pares de bases. En otras realizaciones, el ADNlc o el ADN tumoral circulante (ADNtc) se usa como cadena molde y no es necesaria la fragmentación, dado que el ADNlc o el ADNtc existe en forma de fragmentos cortos. Por ejemplo, el ADNlc fetal circula en el torrente sanguíneo como fragmentos de una longitud de aproximadamente 170 pares de bases (pb) (Fan et al., Clin Chem 56:1279-1286 [2010]), y no se requiere fragmentación del ADN antes de la secuenciación. La tecnología de secuenciación de Illumina se basa en la unión del ADN genómico fragmentado a una superficie plana ópticamente transparente en la que se unen anclajes oligonucleotídicos. Se reparan los extremos de la cadena molde de ADN para generar extremos romos 5'-fosforilados y se utiliza la actividad polimerasa del fragmento Klenow para añadir una única base de A al extremo 3'

de los fragmentos de ADN fosforilados romos. Esta adición prepara los fragmentos de ADN para el ligamiento de adaptadores oligonucleotídicos, que tienen un saliente de una única base de T en su extremo 3' para aumentar la eficiencia del ligamiento. Los oligonucleótidos adaptadores son complementarios a los oligos de anclaje de la célula de flujo. En condiciones de dilución limitante, se añade cadena molde de ADN monocatenario modificado con adaptador a la cubeta de lectura y se inmoviliza mediante hibridación con los oligonucleótidos de anclaje. Se extienden fragmentos de ADN unidos y se amplifican por puente para crear una cubeta de lectura de secuenciación de densidad ultraalta con cientos de millones de grupos, conteniendo cada uno aproximadamente 1000 copias de la misma cadena molde. En una realización, el ADN genómico fragmentado al azar se amplifica usando PCR antes de someterlo a amplificación por clúster. Alternativamente, se usa una preparación de biblioteca genómica libre de amplificación, y el ADN genómico fragmentado al azar se enriquece usando la amplificación de clústeres solo (Kozarewa et al., Nature Methods 6:291-295 [2009]). En algunas aplicaciones, las cadenas molde se secuencian utilizando una robusta tecnología de secuenciación por síntesis de ADN de cuatro colores que emplea terminadores reversibles con colorantes fluorescentes eliminables. La detección de fluorescencia de alta sensibilidad se logra utilizando excitación láser y óptica de reflexión interna total. Las lecturas de secuencia corta que van de aproximadamente decenas a unos cuantos cientos de pares de bases se alinean contra un genoma de referencia y la cartografía única de las lecturas de secuencia corta con el genoma de referencia se identifican usando un software de proceso de análisis de datos especialmente desarrollado. Después de completar la primera lectura, las cadenas molde se pueden regenerar *in situ* para permitir una segunda lectura desde el extremo opuesto de los fragmentos. Por lo tanto, puede usarse una secuenciación de extremo único o de ambos extremos de los fragmentos de ADN.

En diversas realizaciones de la exposición, se puede usar secuenciación por síntesis que permite la secuenciación de ambos extremos. En algunas realizaciones, la plataforma de secuenciación por síntesis de Illumina implica agrupar fragmentos. La agrupación es un proceso en el que cada molécula de fragmento se amplifica isotérmicamente. En algunas realizaciones, como el ejemplo aquí descrito, el fragmento tiene dos adaptadores diferentes unidos a los dos extremos del fragmento, y los adaptadores permiten que el fragmento se hibride con los dos oligos diferentes en la superficie de un carril de cubeta de lectura. El fragmento además incluye o está conectado a dos secuencias de índice en dos extremos del fragmento, secuencias de índice que proporcionan marcadores para identificar diferentes muestras en secuenciación multiplexada. En algunas plataformas de secuenciación, un fragmento que se va a secuenciar a partir de ambos extremos también se denomina inserto.

En alguna implementación, una cubeta de lectura para el agrupamiento en la plataforma de Illumina es un portaobjetos de vidrio con carriles. Cada carril es un canal de vidrio recubierto con un campo de dos tipos de oligos (p. ej., oligos P5 y P7'). La hibridación se habilita mediante el primero de los dos tipos de oligos en la superficie. Este oligo es complementario a un primer adaptador en un extremo del fragmento. Una polimerasa crea una cadena complementaria del fragmento hibridado. La molécula bicatenaria se desnaturaliza y la cadena molde original se elimina por lavado. La cadena restante, en paralelo con muchas otras cadenas restantes, se amplifica clonalmente a través de la amplificación por puente.

En la amplificación por puente y otros métodos de secuenciación que implican agrupamiento, se pliega una cadena y la segunda región adaptadora en un segundo extremo de la cadena se hibrida con el segundo tipo de oligos en la superficie de la cubeta de lectura. Una polimerasa genera una cadena complementaria, formando una molécula de puente bicatenaria. Esta molécula bicatenaria se desnaturaliza dando como resultado dos moléculas monocatenarias unidas a la cubeta de lectura a través de dos oligonucleótidos diferentes. Después, el proceso se repite una y otra vez, y se produce simultáneamente para millones de grupos, dando como resultado la amplificación clonal de todos los fragmentos. Después de la amplificación por puente, las cadenas inversas se escinden y se retiran mediante lavado, dejando solo las cadenas directas. Los extremos 3' se bloquean para impedir el cebado no deseado.

Después del agrupamiento, la secuenciación comienza con la extensión de un primer cebador de secuenciación para generar la primera lectura. Con cada ciclo, los nucleótidos marcados con fluorescencia compiten por la adición a la cadena en crecimiento. Solo se incorpora uno en función de la secuencia de la cadena molde. Después de la adición de cada nucleótido, el grupo se excita con una fuente de luz y se emite una señal fluorescente característica. El número de ciclos determina la longitud de la lectura. La longitud de onda de emisión y la intensidad de señal determinan la llamada de bases. Para un grupo concreto, se leen simultáneamente todas las cadenas idénticas. Se secuencian de manera masiva y en paralelo cientos de millones de grupos. Al finalizar la primera lectura, el producto leído se retira mediante lavado.

En la siguiente etapa de los protocolos que implican dos cebadores de índice, se introduce un cebador de índice 1 y se hibrida con una región índice 1 en la cadena molde. Las regiones de índice proporcionan la identificación de fragmentos, que es útil para demultiplexar muestras en un proceso de secuenciación multiplexada. La lectura del índice 1 se genera de un modo similar a la primera lectura. Después de completar la lectura del índice 1, el producto de lectura se elimina por lavado y se desprotege el extremo 3' de la cadena. A continuación, se pliega la cadena molde y se une a un segundo oligonucleótido en la cubeta de lectura. Una secuencia de índice 2 se lee de la misma manera que una de índice 1. A continuación, un producto de lectura de índice 2 se retira mediante lavado al finalizar la etapa.

Después de leer dos índices, la lectura 2 se inicia usando polimerasas para extender los segundos oligonucleótidos de las cubetas de lectura, formando un puente bicatenario. Este ADN bicatenario se desnaturaliza y se bloquea el

extremo 3'. La cadena directa original se escinde y se retira mediante lavado, dejando la cadena inversa. La lectura 2 comienza con la introducción de un cebador de secuenciación de lectura 2. Al igual que con la lectura 1, las etapas de secuenciación se repiten hasta que se logra la longitud deseada. El producto de lectura 2 se retira mediante lavado. Este proceso completo genera millones de lecturas, que representan todos los fragmentos. Las secuencias de las bibliotecas de muestras agrupadas se separan en función de los índices únicos introducidos durante la preparación de la muestra. Para cada muestra, las lecturas de tramos similares de llamadas de base se agrupan localmente. Las lecturas directas e inversas se emparejan creando secuencias contiguas. Estas secuencias contiguas se alinean con el genoma de referencia para la identificación de la variante.

El ejemplo de secuenciación por síntesis arriba descrito implica lecturas de ambos extremos, lo que se usa en muchas de las realizaciones de los métodos expuestos. La secuenciación de ambos extremos implica 2 lecturas de los dos extremos de un fragmento. Las lecturas de ambos extremos se usan para resolver alineamientos ambiguos. La secuenciación de ambos extremos permite a los usuarios elegir la longitud del inserto (o el fragmento que se va a secuenciar) y secuenciar cualquiera de los extremos del inserto, generando datos de secuencia alineables de alta calidad. Debido a que se conoce la distancia entre cada lectura emparejada, los algoritmos de alineamiento pueden usar esta información para cartografiar lecturas sobre regiones repetitivas con mayor precisión. Esto da lugar a un mejor alineamiento de las lecturas, especialmente a través de regiones repetitivas difíciles de secuenciar del genoma. La secuenciación de ambos extremos puede detectar reordenamientos, incluidas inserciones y deleciones (indel) e inversiones.

Las lecturas de ambos extremos pueden usar insertos de diferente longitud (es decir, un tamaño de fragmento diferente a secuenciar). Como significado predeterminado en esta exposición, las lecturas de ambos extremos se usan para referirse a las lecturas obtenidas a partir de diversas longitudes de inserto. En algunos casos, para distinguir las lecturas de ambos extremos de insertos cortos de las lecturas de ambos extremos de insertos largos, estas últimas se denominan específicamente lecturas de pares de parejas. En algunas realizaciones que implican lecturas de pares de parejas, dos adaptadores de unión a biotina se unen primero a dos extremos de un inserto relativamente largo (p. ej., de varios kb). Los adaptadores de unión a biotina unen a continuación los dos extremos del inserto, formando una molécula circularizada. A continuación, puede obtenerse un subfragmento que abarca los adaptadores de unión a biotina fragmentando adicionalmente la molécula circularizada. El subfragmento que incluye los dos extremos del fragmento original en orden de secuencia opuesto puede secuenciarse a continuación mediante el mismo procedimiento que para la secuenciación de ambos extremos de inserto corto descrita anteriormente. Encontrará detalles adicionales sobre la secuenciación de pares de parejas mediante una plataforma de Illumina en una publicación en línea en la siguiente dirección, que se menciona como referencia en su totalidad: res.illumina.com/documents/products/technotes/technote_nextera_matepair_data_pro cessing.pdf

Tras la secuenciación de los fragmentos de ADN, las lecturas de secuencia de una longitud predeterminada, p. ej., 100 pb, se localizan cartografiándolas (alineándolas) con un genoma de referencia conocido. Las lecturas cartografiadas y sus ubicaciones correspondientes en la secuencia de referencia también se denominan marcadores. En otra realización del procedimiento, la localización se realiza compartiendo k-meros y alineando una lectura con otra. Los análisis de muchas realizaciones divulgadas en la presente memoria emplean lecturas que o bien están mal alineadas o que no pueden alinearse, así como lecturas alineadas (marcadores). En una realización, la secuencia del genoma de referencia es la secuencia NCBI36/hg18, que está disponible en Internet en genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105. Como alternativa, la secuencia del genoma de referencia es GRCh37/hg19 o GRCh38, que se encuentra disponible en Internet en genome.ucsc.edu/cgi-bin/hgGateway. Otras fuentes de información de secuencia públicas incluyen GenBank, dbEST, dbSTS, EMBL (el Laboratorio Europeo de Biología Molecular), y DDBJ (la base de datos de ADN de Japón). Hay disponibles varios algoritmos informáticos para alinear secuencias, incluidos, sin limitarse a ellos, BLAST (Altschul et al., 1990), BLITZ (MPsrch) (Sturrock & Collins, 1993), FASTA (Person & Lipman, 1988), BOWTIE (Langmead et al., Genome Biology 10:R25.1-R25.10 [2009]), o ELAND (Illumina, Inc., San Diego, CA, EE.UU.). En una realización, un extremo de las copias clonalmente expandidas de las moléculas de ADNlc plasmático se secuencia y procesa mediante análisis de alineamiento bioinformático para Illumina Genome Analyzer (analizador de genoma de Illumina), que usa el software Efficient Large-Scale Alignment of Nucleotide Databases (ELAND, Alineación eficiente a gran escala de bases de datos de nucleótidos).

También se pueden usar otros métodos de secuenciación para obtener lecturas de secuencia y alineaciones de las mismas. Métodos adecuados adicionales se describen en la solicitud de patente US-15/130,668 presentada el 15 de abril de 2016, que se menciona como referencia en su totalidad.

En algunas realizaciones de los métodos descritos en la presente memoria, las lecturas de secuencia tienen aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130 pb, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb, o aproximadamente 500 pb. Se espera que los avances tecnológicos permitan lecturas de un solo extremo de más de

500 pb permitiendo lecturas de más de aproximadamente 1000 pb cuando se generen lecturas de ambos extremos. En algunas realizaciones, las lecturas de extremos apareados se usan para determinar secuencias de interés, que incluyen lecturas de secuencia que son de aproximadamente 20 pb a 1000 pb, de aproximadamente 50 pb a 500 pb o de 80 pb a 150 pb. En diversas realizaciones, las lecturas de ambos extremos se utilizan para evaluar una secuencia de interés. La secuencia de interés es más larga que las lecturas. En algunas realizaciones, la secuencia de interés tiene una longitud mayor que aproximadamente 100 pb, 500 pb, 1000 pb o 4000 pb. La cartografía de las lecturas de secuencia se logra comparando la secuencia de las lecturas con la secuencia de la referencia para determinar el origen cromosómico de la molécula de ácido nucleico secuenciada, y no se necesita información de secuencia genética específica. Se puede permitir que se dé un pequeño grado de emparejamientos erróneos (0–2 emparejamientos erróneos por lectura) para polimorfismos minoritarios que pueden existir entre el genoma de referencia y los genomas en la muestra mixta. En algunas realizaciones, se usan lecturas que están alineadas con la secuencia de referencia como lecturas de anclaje, y se usan lecturas emparejadas a las lecturas de anclaje pero que no pueden alinearse o se alinean mal a la referencia como lecturas ancladas. En algunas realizaciones, las lecturas mal alineadas pueden tener un número relativamente grande de porcentaje de emparejamientos erróneos por lectura, p. ej., al menos aproximadamente el 5 %, al menos aproximadamente el 10 %, al menos aproximadamente el 15 %, o al menos aproximadamente el 20 % de emparejamientos erróneos por lectura.

De forma típica, se obtiene una pluralidad de marcadores de secuencia (es decir, lecturas alineadas a una secuencia de referencia) por muestra. En algunas realizaciones, al menos aproximadamente 3×10^6 marcadores de secuencia, al menos aproximadamente 5×10^6 marcadores de secuencia, al menos aproximadamente 8×10^6 marcadores de secuencia, al menos aproximadamente 10×10^6 marcadores de secuencia, al menos aproximadamente 15×10^6 marcadores de secuencia, al menos aproximadamente 20×10^6 marcadores de secuencia, al menos aproximadamente 30×10^6 marcadores de secuencia, al menos aproximadamente 40×10^6 marcadores de secuencia, o al menos aproximadamente 50×10^6 marcadores de secuencia de, p. ej., 100 pb, se obtienen de las lecturas que se cartografían con el genoma de referencia por muestra. En algunas realizaciones, todas las lecturas de secuencia se cartografían a todas las regiones del genoma de referencia, proporcionando lecturas de todo el genoma. En otras realizaciones, las lecturas se cartografían a una secuencia de interés.

Aparatos y sistemas para secuenciación que utilizan UMI

Como debería ser evidente, determinadas realizaciones de la invención emplean procesos que actúan bajo el control de instrucciones y/o datos almacenados en uno o más sistemas informáticos o transferidos a través de ellos. Ciertas realizaciones también se refieren a un aparato para realizar estas operaciones. Este aparato puede diseñarse y/o construirse especialmente para los fines requeridos, o puede ser un ordenador de uso general configurado selectivamente por uno o más programas informáticos y/o estructuras de datos almacenados o puestos a disposición del ordenador de cualquier otra manera. En particular, se pueden usar diversas máquinas de uso general con programas escritos según las enseñanzas de la presente memoria, o puede ser más conveniente construir un aparato más especializado para llevar a cabo las etapas del método requeridas. A continuación, se muestra y describe una estructura particular para una variedad de estas máquinas.

Ciertas realizaciones también proporcionan funcionalidad (por ejemplo, código y procesos) para almacenar cualquiera de los resultados (por ejemplo, resultados de consultas) o estructuras de datos generadas como se describe en la presente descripción. Tales resultados o estructuras de datos se almacenan típicamente, al menos temporalmente, en un medio legible por ordenador. Los resultados o las estructuras de datos también pueden emitirse de diversas maneras, tales como visualización, impresión y similares.

Ejemplos de medios tangibles legibles por ordenador adecuados para usar en productos de programas informáticos y aparatos computacionales de esta invención incluyen, aunque no de forma limitativa, medios magnéticos tales como discos duros, disquetes y cintas magnéticas; medios ópticos tales como discos CD-ROM; medios magneto-ópticos; dispositivos de memoria semiconductora (p. ej., memoria flash) y dispositivos de hardware que están especialmente configurados para almacenar y ejecutar instrucciones de programación, tales como dispositivos de memoria de solo lectura (ROM) y memoria de acceso aleatorio (RAM) y, a veces, circuitos integrados específicos de aplicaciones (ASIC), dispositivos lógicos programables (PLD) y medios de transmisión de señales para entregar instrucciones legibles por ordenador, tales como redes de área local, redes de área amplia e Internet. Los datos y las instrucciones de programa proporcionados en la presente descripción también pueden incorporarse en una onda portadora u otro medio de transporte (incluidas las vías electrónicas u ópticamente conductoras). Los datos y las instrucciones del programa de esta invención también pueden estar incorporadas en una onda portadora u otro medio de transporte (p. ej., líneas ópticas, líneas eléctricas y/u ondas).

Los ejemplos de instrucciones de programa incluyen el código de bajo nivel, tal como el producido por un compilador, así como código de nivel superior, que puede ser ejecutado por el ordenador usando un intérprete. Además, las instrucciones del programa pueden ser código máquina, código fuente y/o cualquier otro código que controle directa o indirectamente el funcionamiento de una máquina informática. El código puede especificar la entrada, la salida, los cálculos, los condicionales, las ramas, los bucles iterativos, etc.

El análisis de los datos de secuenciación y el diagnóstico derivado de los mismos se realizan de forma típica usando diversos algoritmos y programas ejecutados por ordenador. Por lo tanto, determinadas realizaciones emplean procesos que implican datos almacenados en o transferidos a través de uno o más sistemas informáticos u otros sistemas de procesamiento. Las realizaciones expuestas en la presente memoria también se refieren al aparato para realizar estas operaciones. Este aparato puede construirse especialmente para los fines requeridos, o puede ser un ordenador de propósito general (o un grupo de ordenadores) activado o reconfigurado selectivamente por un programa informático y/o una estructura de datos almacenada en el ordenador. En algunas realizaciones, un grupo de procesadores realiza algunas o todas las operaciones analíticas enumeradas colaborativamente (p. ej., a través de una red o informática en la nube) y/o en paralelo. Un procesador o grupo de procesadores para realizar los métodos descritos en la presente memoria puede ser de diversos tipos, incluidos microcontroladores y microprocesadores, tales como dispositivos programables (p. ej., CPLD y FPGA) y dispositivos no programables, tales como ASIC de matriz de puertas o microprocesadores de propósito general.

Una implementación proporciona un sistema para su uso en la determinación de una secuencia con baja frecuencia alélica en una muestra de ensayo que incluye ácidos nucleicos, incluyendo el sistema un secuenciador para recibir una muestra de ácido nucleico y proporcionar información de secuencia de ácido nucleico a partir de la muestra; un procesador; y un medio de almacenamiento legible por ordenador que tiene almacenadas en él instrucciones de ejecución en dicho procesador para determinar una secuencia de interés en la muestra de ensayo mediante lo siguiente: (a) aplicar adaptadores a los fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador comprende un índice molecular único no aleatorio y en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos no aleatorios y de longitud variable (vNRUMI); (b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados; (c) secuenciar, mediante el secuenciador, la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI; (d) identificar, mediante el procesador, entre la pluralidad de lecturas, las lecturas asociadas con un mismo índice molecular único no aleatorio de longitud variable (vNRUMI); y (e) determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI.

En algunas realizaciones de cualquiera de los sistemas proporcionados en la presente memoria, el secuenciador está configurado para realizar la secuenciación de próxima generación (NGS). En algunas realizaciones, el secuenciador está configurado para realizar una secuenciación masivamente en paralelo usando secuenciación por síntesis con terminadores de colorante reversibles. En otras realizaciones, el secuenciador está configurado para realizar la secuenciación por ligadura. En otras realizaciones adicionales, el secuenciador está configurado para realizar secuenciación de una sola molécula.

Otra implementación proporciona un sistema que incluye un sintetizador de ácidos nucleicos, un procesador y un medio de almacenamiento legible por máquina que tiene almacenadas instrucciones para su ejecución en dicho procesador para preparar los adaptadores de secuenciación. Las instrucciones incluyen: (a) proporcionar, por parte del procesador, un conjunto de secuencias de oligonucleótidos que tengan al menos dos longitudes moleculares diferentes; (b) seleccionar, por parte del procesador, un subconjunto de secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos, todas las distancias de modificación entre las secuencias de oligonucleótidos del subconjunto de secuencias de oligonucleótidos que cumplen un valor umbral, el subconjunto de secuencias de oligonucleótidos que forman un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI); y (c) sintetizar, mediante el sintetizador de ácido nucleico, una pluralidad de adaptadores de secuenciación, en donde cada adaptador de secuenciación comprende una región hibridada bicatenaria, un brazo 5' monocatenario, un brazo 3' monocatenario y al menos un vNRUMI del conjunto de vNRUMI.

Además, determinadas realizaciones se refieren a medios legibles por ordenador tangibles y/o no transitorios o productos de programa informático que incluyen instrucciones de programa y/o datos (incluidas estructuras de datos) para realizar diversas operaciones implementadas por ordenador. Los ejemplos de medios legibles por ordenador incluyen, aunque no de forma limitativa, dispositivos de memoria de semiconductores, medios magnéticos tales como unidades de disco, cinta magnética, medios ópticos tales como CD, medios magnetoópticos y dispositivos de hardware que están configurados especialmente para almacenar y ejecutar instrucciones de programa, tales como dispositivos de memoria de solo lectura (ROM) y memoria de acceso aleatorio (RAM). Los medios legibles por ordenador pueden ser controlados directamente por un usuario final o los medios pueden ser controlados indirectamente por el usuario final. Los ejemplos de medios controlados directamente incluyen los medios ubicados en una instalación de usuario y/o medios que no se comparten con otras entidades. Los ejemplos de medios controlados indirectamente incluyen medios que son indirectamente accesibles para el usuario a través de una red externa y/o mediante un servicio que proporciona recursos compartidos tales como la "nube". Los ejemplos de instrucciones de programa incluyen tanto código de máquina, tales como las producidas por un compilador, y archivos que contienen código de nivel superior que puede ejecutarse por el ordenador usando un intérprete.

En diversas realizaciones, los datos o información empleada en los métodos y aparatos expuestos se proporcionan en un formato electrónico. Tales datos o información pueden incluir lecturas y marcadores derivadas de una muestra de ácido nucleico, secuencias de referencia (incluidas secuencias de referencia que proporcionan única o principalmente polimorfismos), llamadas, tales como llamadas de diagnóstico de cáncer, recomendaciones de

asesoramiento, diagnósticos y similares. Como se utiliza en la presente memoria, los datos u otra información proporcionada en formato electrónico está disponible para su almacenamiento en una máquina y transmisión entre máquinas. Convencionalmente, los datos en formato electrónico se proporcionan digitalmente y pueden almacenarse como bits y/o bytes en diversas estructuras de datos, listas, bases de datos, etc. Los datos pueden incorporarse electrónicamente, ópticamente, etc.

Una realización proporciona un producto de programa informático para generar una salida que indica la secuencia de un fragmento de ADN de interés en una muestra de ensayo. El producto informático puede contener instrucciones para realizar uno cualquiera o más de los métodos anteriormente descritos para determinar una secuencia de interés. Como ya se ha explicado, el producto informático puede incluir un medio legible por ordenador no transitorio y/o tangible que tiene una lógica ejecutable por ordenador o compilable (p. ej., instrucciones) registrada sobre el mismo para permitir que un procesador determine una secuencia de interés. En un ejemplo, el producto informático incluye un medio legible por ordenador que tiene una lógica ejecutable o compilable por ordenador (p. ej., instrucciones) grabada en el mismo para permitir a un procesador diagnosticar una afección o determinar una secuencia de ácido nucleico de interés.

Debe entenderse que no es práctico, o incluso posible en la mayoría de los casos, que un ser humano sin ayuda realice las operaciones computacionales de los métodos expuestos en la presente memoria. Por ejemplo, la correlación de una única lectura de 30 pb a partir de una muestra a uno cualquiera de los cromosomas humanos puede requerir años de esfuerzo sin la ayuda de un aparato informático. Por supuesto, el problema se hace más complejo ya que las llamadas fiables de mutaciones de baja frecuencia alélica generalmente requieren cartografiar miles (p. ej., al menos aproximadamente 10.000) o incluso millones de lecturas en uno o más cromosomas.

Los métodos divulgados en la presente memoria pueden realizarse empleando un sistema para determinar una secuencia de interés en una muestra de ensayo. El sistema puede incluir: (a) un secuenciador para recibir ácidos nucleicos de la muestra de prueba que proporciona información de secuencia de ácidos nucleicos de la muestra; (b) un procesador; y (c) uno o más medios legibles por ordenador que tienen almacenados en los mismos instrucciones ejecutables en dicho procesador para determinar una secuencia de interés en la muestra de ensayo. En algunas realizaciones, los métodos se instruyen por un medio legible por ordenador que tiene almacenadas en el mismo instrucciones legibles por ordenador para llevar a cabo un método para determinar la secuencia de interés. Por tanto, una realización proporciona un producto de programa informático que incluye un medio legible por ordenador no transitorio que almacena código de programación que, cuando se ejecuta por uno o más procesadores de un sistema informático, hace que el sistema informático implemente un método para determinar las secuencias de fragmentos de ácido nucleico en una muestra de ensayo. El código de programa puede incluir: (a) código para obtener una pluralidad de lecturas de una pluralidad de polinucleótidos amplificados, cada polinucleótido de la pluralidad de polinucleótidos amplificados que comprende un adaptador unido a un fragmento de ADN, en donde el adaptador comprende un índice molecular único no aleatorio, y en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes, formando un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI); (b) código para identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo VNRUMI; y (c) código para determinar, mediante las lecturas asociadas con el mismo vNRUMI, una secuencia de un fragmento de ADN de la muestra.

En algunas realizaciones, los códigos del programa las instrucciones pueden incluir además el registro automático de información pertinente al método. El registro médico del paciente puede ser mantenido, por ejemplo, por un laboratorio, consultorio médico, hospital, organización de mantenimiento de la salud, compañía de seguros o un sitio web de registro médico personal. Además, basándose en los resultados del análisis implementado por procesador, el método puede implicar además prescribir, iniciar y/o alterar el tratamiento de un sujeto humano del que se ha tomado la muestra de prueba. Esto puede implicar realizar una o más pruebas o análisis adicionales en muestras adicionales tomadas del sujeto.

Los métodos divulgados también pueden realizarse utilizando un sistema de procesamiento informático que está adaptado o configurado para realizar un método para determinar una secuencia de interés. Una realización proporciona un sistema de procesamiento informático que está adaptado o configurado para realizar un método como se describe en la presente memoria. En una realización, el aparato incluye un dispositivo de secuenciación adaptado o configurado para secuenciar al menos una porción de las moléculas de ácido nucleico de una muestra para obtener el tipo de información de secuencia descrita en cualquier otro sitio de la presente memoria. El aparato también puede incluir componentes para procesar la muestra. Tales componentes se describen en cualquier otro sitio de la presente memoria.

La secuencia u otros datos pueden introducirse en un ordenador o almacenarse en un medio legible por ordenador, ya sea directa o indirectamente. En una realización, un sistema informático se acopla directamente a un dispositivo de secuenciación que lee y/o analiza secuencias de ácidos nucleicos de muestras. Las secuencias u otra información obtenida de tales herramientas se proporcionan al sistema informático mediante una interfaz. Como alternativa, las secuencias procesadas por el sistema se proporcionan desde una fuente de almacenamiento de secuencia, tal como una base de datos u otro repositorio. Una vez disponible para el aparato de procesamiento, un dispositivo de memoria o un dispositivo de almacenamiento masivo recoge o almacena, al menos temporalmente, secuencias de los ácidos

nucleicos. Además, el dispositivo de memoria puede almacenar recuentos de marcadores para diversos cromosomas o genomas, etc. La memoria también puede almacenar diversas rutinas y/o programas para analizar la presentación de la secuencia o los datos cartografiados. Tales programas/rutinas pueden incluir programas para realizar análisis estadísticos, etc.

En un ejemplo, un usuario proporciona una muestra en un aparato de secuenciación. Los datos se recopilan y/o analizan mediante el aparato de secuenciación que está conectado a un ordenador. El software en el ordenador permite la recopilación y/o el análisis de datos. Los datos pueden almacenarse, mostrarse (a través de un monitor u otro dispositivo similar), y/o enviarse a otra ubicación. El ordenador puede estar conectado a internet, que se usa para transmitir datos a un dispositivo portátil utilizado por un usuario remoto (p. ej., un médico, científico o analista). Se entiende que los datos pueden almacenarse y/o analizarse antes de transmitirlos. En algunas realizaciones, los datos sin procesar se recopilan y envían a un usuario o aparato remoto que analizará y/o almacenará los datos. La transmisión puede realizarse a través de internet, pero también puede producirse mediante satélite u otra conexión. Como alternativa, los datos pueden almacenarse en un medio legible por ordenador y el medio puede enviarse a un usuario final (p. ej., mediante correo). El usuario remoto puede estar en la misma ubicación geográfica o diferente, incluidos, aunque no de forma limitativa, un edificio, ciudad, estado, país o continente.

En algunas realizaciones, los métodos también incluyen recopilar datos con respecto a una pluralidad de secuencias de polinucleótidos (p. ej., lecturas, marcadores y/o secuencias cromosómicas de referencia) y enviar los datos a un ordenador u otro sistema informático. Por ejemplo, el ordenador puede estar conectado a un equipo de laboratorio, p. ej., un aparato de recolección de muestras, un aparato de amplificación de nucleótidos, un aparato de secuenciación de nucleótidos o un aparato de hibridación. El ordenador puede entonces recopilar datos aplicables recopilados por el dispositivo de laboratorio. Los datos pueden almacenarse en un ordenador en cualquier etapa, p. ej., mientras se recopilan en tiempo real, antes del envío, durante o junto con el envío, o siguiendo el envío. Los datos pueden almacenarse en un medio legible por ordenador que puede extraerse del ordenador. Los datos recogidos o almacenados pueden transmitirse desde el ordenador hasta una ubicación remota, p. ej., a través de una red local o una red de área amplia tal como internet. En la ubicación remota, los datos transmitidos se pueden someter a diversas operaciones, como se describe a continuación.

Entre los tipos de datos con formato electrónico que pueden almacenarse, transmitirse, analizarse y/o manipularse en los sistemas, aparatos y métodos expuestos en la presente memoria se hallan los siguientes:

Lecturas obtenidas mediante secuenciación de ácidos nucleicos en una muestra de prueba

Marcadores obtenidos alineando las lecturas con un genoma de referencia u otra secuencia o secuencias de referencia

El genoma o secuencia de referencia

Umbrales para llamar a una muestra de prueba, ya sea afectada, no afectada o sin llamada

Las llamadas reales de las afecciones médicas relacionadas con la secuencia de interés

Diagnóstico (condición clínica asociada a las llamadas)

Recomendaciones para pruebas adicionales derivadas de las llamadas y/o diagnósticos

Tratamiento y/o planes de monitorización derivados de las llamadas y/o diagnósticos

Estos diversos tipos de datos pueden obtenerse, almacenarse, transmitirse, analizarse y/o manipularse en una o más ubicaciones usando un aparato distinto. Las opciones de procesamiento abarcan un amplio espectro. En un extremo del espectro, toda o gran parte de esta información se almacena y se usa en la ubicación donde la muestra de prueba se procesa, p. ej., una consulta médica u otro entorno clínico. En otro extremo, la muestra se obtiene en una ubicación, se procesa y se secuencia opcionalmente en una ubicación diferente, las lecturas se alinean y las llamadas se realizan en una o más ubicaciones diferentes, y los diagnósticos, recomendaciones y/o planes se preparan en otra ubicación (que puede ser una ubicación donde se obtuvo la muestra).

En diversas realizaciones, las lecturas se generan con el aparato de secuenciación y después se transmiten a un punto remoto, donde se procesan para determinar una secuencia de interés. En esta ubicación remota, a modo de ejemplo, las lecturas se alinean con una secuencia de referencia para producir lecturas de anclaje y ancladas. Entre las operaciones de procesamiento que pueden emplearse en diferentes ubicaciones se hallan las siguientes:

Recolección de muestras

Procesamiento de muestra preliminar a la secuenciación

Secuenciación

Análisis de datos de secuencia y obtención de llamadas médicas

Diagnóstico

Informe acerca de un diagnóstico y/o una llamada al paciente o proveedor de atención sanitaria

Desarrollo de un plan para tratamiento, prueba y/o monitorización adicionales

Ejecución del plan

Asesoramiento

Una cualquiera o más de estas operaciones pueden automatizarse como se describe en cualquier otro sitio de la presente memoria. Normalmente, la secuenciación y el análisis de datos de secuencia y la obtención de llamadas médicas se realizarán por medios informáticos. Las otras operaciones pueden realizarse manual o automáticamente.

La Figura 6 muestra una implementación de un sistema disperso para producir una llamada o diagnóstico a partir de una muestra de prueba. Se usa una ubicación de recolección de muestras 01 para obtener una muestra de prueba de un paciente. Las muestras se proporcionaron a continuación a una ubicación de procesamiento y secuenciación 03 donde la muestra de prueba puede procesarse y secuenciarse como se ha descrito anteriormente. La ubicación 03 incluye un aparato para procesar la muestra así como un aparato para secuenciar la muestra procesada. El resultado de la secuenciación, como se describe en cualquier otro sitio de la presente memoria, es una colección de lecturas que típicamente se proporcionan en un formato electrónico y se proporcionan a una red tal como Internet, lo que se indica por el número de referencia 05 en la Figura 6.

Los datos de secuencia se proporcionan a una ubicación remota 07 donde se realizan el análisis y la generación de llamadas. Esta ubicación puede incluir uno o más dispositivos informáticos potentes tales como ordenadores o procesadores. Una vez que los recursos computacionales en la ubicación 07 han completado su análisis y generado una llamada desde la información de secuencia recibida, la llamada se retransmite a la red 05. En algunas implementaciones, no solo se genera una llamada en la ubicación 07, sino que también se genera un diagnóstico asociado. La llamada y/o el diagnóstico se transmiten a continuación a través de la red y de vuelta a la ubicación de recolección de muestras 01 como se ilustra en la Figura 6. Como se ha explicado, esta es simplemente una de muchas variaciones de cómo las diversas operaciones asociadas con la generación de una llamada o diagnóstico pueden dividirse entre diversas ubicaciones. Una variante común implica proporcionar recolección y procesamiento de muestras y secuenciación en una única ubicación. Otra variación implica proporcionar procesamiento y secuenciación en la misma ubicación que el análisis y la generación de llamadas.

La Figura 7 ilustra, en un formato de bloque simple, un sistema informático típico que, cuando se configura o diseña adecuadamente, puede servir como aparato computacional de acuerdo con determinadas realizaciones. El sistema informático 2000 incluye cualquier número de procesadores 2002 (también denominados unidades centrales de procesamiento o CPU) que están acoplados a dispositivos de almacenamiento, incluido el almacenamiento primario 2006 (normalmente una memoria de acceso aleatorio o RAM), el almacenamiento primario 2004 (normalmente una memoria de solo lectura o ROM). La CPU 2002 puede ser de diversos tipos, incluidos microcontroladores y microprocesadores, tales como dispositivos programables (p. ej., CPLD y FPGA) y dispositivos no programables, tales como ASIC de matriz de puertas o microprocesadores de propósito general. En la realización representada, el almacenamiento primario 2004 actúa para transferir datos e instrucciones unidireccionalmente a la CPU y el almacenamiento primario 2006 se usa normalmente para transferir datos e instrucciones de manera bidireccional. Ambos dispositivos de almacenamiento primario pueden incluir cualquier medio legible por ordenador adecuado, tal como los descritos anteriormente. Un dispositivo de almacenamiento masivo 2008 también está acoplado bidireccionalmente al almacenamiento primario 2006 y proporciona una capacidad de almacenamiento de datos adicional y puede incluir cualquiera de los medios legibles por ordenador descritos anteriormente. El dispositivo de almacenamiento masivo 2008 se puede usar para almacenar programas, datos y similares y normalmente es un medio de almacenamiento secundario, tal como un disco duro. Frecuentemente, tales programas, datos y similares se copian temporalmente a la memoria primaria 2006 para su ejecución en la CPU 2002. Se apreciará que la información retenida dentro del dispositivo de almacenamiento masivo 2008 puede, en los casos apropiados, incorporarse de manera estándar como parte del almacenamiento primario 2004. Un dispositivo de almacenamiento masivo específico, tal como un CD-ROM 2014, también puede pasar datos de forma unidireccional a la CPU o al almacenamiento primario.

La CPU 2002 también está acoplada a una interfaz 2010 que se conecta a uno o más dispositivos de entrada/salida, tales como un secuenciador de ácidos nucleicos (2020), un sintetizador de ácidos nucleicos (2022), monitores de vídeo, bolas de seguimiento, ratones, teclados, micrófonos, pantallas sensibles al tacto, lectores de tarjetas transductoras, lectores de cintas magnéticas o de papel, tabletas, lápices ópticos, periféricos de reconocimiento de voz o escritura a mano, puertos USB u otras entradas conocidas dispositivos como, por supuesto, otros ordenadores. Finalmente, la CPU 2002 puede acoplarse opcionalmente a un dispositivo externo, tal como una base de datos o una red informática o de telecomunicaciones, usando una conexión externa, como se muestra generalmente en 2012. Con

una conexión de este tipo, se contempla que la CPU pueda recibir información de la red o enviar información a la red durante la realización de las etapas del método descritas en la presente descripción. En algunas implementaciones, un secuenciador de ácidos nucleicos o un sintetizador de ácidos nucleicos se pueden vincular comunicativamente a la CPU 2002 a través de la conexión de red 2012 en lugar de o además de a través de la interfaz 2010.

En una realización, un sistema tal como el sistema informático 2000 se usa como un sistema de importación, correlación de datos y consulta capaz de realizar algunas o todas las tareas descritas en la presente memoria. La información y los programas, incluidos los archivos de datos, pueden proporcionarse a través de una conexión de red 2012 para que un investigador acceda a ellos o los descargue. Alternativamente, tal información, programas y archivos pueden proporcionarse al investigador en un dispositivo de almacenamiento.

En una realización específica, el sistema informático 2000 está acoplado directamente a un sistema de adquisición de datos, tal como una micromatriz, un sistema de cribado de alto rendimiento o un secuenciador de ácidos nucleicos (2020) que captura datos de muestras. Los datos de tales sistemas se proporcionan a través de la interfaz 2010 para que el sistema 2000 los analice. Como alternativa, los datos procesados por el sistema 2000 se proporcionan desde una fuente de almacenamiento de datos, tal como una base de datos u otro repositorio de datos relevantes. Una vez en el aparato 2000, un dispositivo de memoria, tal como el almacenamiento primario 2006 o el almacenamiento masivo 2008, guarda o almacena, al menos temporalmente, los datos relevantes. La memoria también puede almacenar diversas rutinas y/o programas para importar, analizar y presentar los datos, incluidas lecturas de secuencias, UMI, códigos para determinar lecturas de secuencias, agrupar lecturas de secuencias y corregir errores en las lecturas, etc.

En ciertas realizaciones, los ordenadores usados en la presente memoria pueden incluir un terminal de usuario, que puede ser cualquier tipo de ordenador (p. ej., ordenador de sobremesa, portátil, tableta, etc.), plataformas de computación multimedia (p. ej., decodificadores por cable, satélite, grabadoras de vídeo digitales, etc.), dispositivos informáticos portátiles (p. ej., PDA, clientes de correo electrónico, etc.), teléfonos móviles o cualquier otro tipo de plataforma informática o de comunicación.

En ciertas realizaciones, los ordenadores usados en la presente descripción también pueden incluir un sistema servidor en comunicación con un terminal de usuario, cuyo sistema servidor puede incluir un dispositivo servidor o dispositivos servidores descentralizados, y puede incluir ordenadores centrales, miniordenadores, superordenadores, ordenadores personales o combinaciones de los mismos. También se puede usar una pluralidad de sistemas de servidores sin apartarse del alcance de la presente invención. Los terminales de usuario y un sistema servidor pueden comunicarse entre sí a través de una red. La red puede comprender, p. ej., redes cableadas tales como LAN (redes de área local), WAN (redes de área amplia), MAN (redes de área metropolitana), ISDN (redes digitales de servicios integrados), etc., así como redes inalámbricas tales como LAN inalámbricas, CDMA, Bluetooth y redes de comunicación por satélite, etc., sin limitar el alcance de la presente invención.

Experimentos

Ejemplo 1

Comparación del método de vNRUMI y otros métodos de códigos de barras

La Tabla 1 muestra la heterogeneidad de pares de bases de los NRUMI, en comparación con la heterogeneidad de pares de bases de los vNRUMI de acuerdo con algunas implementaciones. Este conjunto de 120 vNRUMI comprende 50 hexámeros y 70 heptámeros. El conjunto de NRUMI se comprende en su totalidad de 218 hexámeros, en los que la distancia mínima de modificación entre dos NRUMI cualesquiera supera un valor umbral. La Tabla 1 presupone que cada uno de los 218 o 128 códigos de barras estaba presente en cantidades iguales, p. ej., hay 1000 de cada UMI. Para la 7.ª base, el nuevo conjunto de vNRUMI tiene una heterogeneidad mucho mejor que el conjunto de NRUMI original, y supera con creces el mínimo recomendado de composición del 5 % por base. Por lo tanto, está claro que el diseño de vNRUMI aborda el desafío mencionado anteriormente de la falta de diversidad de pares de bases en determinados ciclos. Otros conjuntos de códigos de barras comprendidos exclusivamente de hexámeros tienen una heterogeneidad por base similar a la del conjunto de NRUMI original que se muestra a continuación.

Tabla 1: Heterogeneidad de pares de bases dentro de las posiciones de UMI

	NRUMI (n = 218)				*vNRUMI (n = 120)*			
Base	A	C	G	T	A	C	G	T
1	0,2431	0,2523	0,1972	0,3073	0,2667	0,2333	0,2417	0,2583
2	0,2844	0,2844	0,1468	0,2844	0,2500	0,2583	0,2250	0,2667
3	0,2431	0,2385	0,2523	0,2661	0,3083	0,2000	0,2500	0,2417
4	0,2110	0,2936	0,1514	0,3440	0,2583	0,2500	0,2750	0,2167

	NRUMI (n = 218)				*vNRUMI (n = 120)*			
5	0,2018	0,2248	0,4083	0,1651	0,3000	0,1833	0,2167	0,3000
6	0,2018	0,3302	0,1009	0,3670	0,2750	0,2750	0,2667	0,1833
7	0	0	0	1	0,1917	0,1750	0,2167	0,4167

Utilizando los NRUMI y vNRUMI anteriores, se realizaron ensayos de simulación *in silicio* para simular 10 000 códigos de barras, se mutó cada código de barras mediante la mutación de cada base de forma independiente, y se intentó recuperar la secuencia de UMI original. La simulación utilizó una tasa de mutación del 2 % en cada base (1 % de probabilidad para SNV, 1 % de probabilidad para indel de tamaño 1). Tenga en cuenta que esta tasa de mutación es apreciablemente superior a las tasas de error de secuenciación típicas de Illumina. Cada una de las 10 000 simulaciones contenía al menos una mutación.

Para proporcionar una comparación adicional con otros métodos que utilizan UMI, en este ensayo de simulación también se utiliza un conjunto de 114 secuencias de NRUMI de longitud 6 nt generadas de acuerdo con un planteamiento existente nxCode. Consulte http://hannonlab.cshl.edu/nxCode/nxCode/main.html. Estas secuencias se sometieron al mismo proceso de mutación descrito anteriormente. El planteamiento nxCode utiliza un modelo probabilístico para determinar mutaciones, y utiliza un planteamiento semiambicioso para obtener un conjunto de NRUMI que tengan la misma longitud molecular. Los resultados de la comparación entre los conjuntos de vNRUMI, NRUMI y nxCode se pueden encontrar en la Tabla 2.

Tabla 2: Resultados de referencia que comparan las tasas de corrección de errores de diferentes diseños de UMI

Métrica	**vNRUMI**	**NRUMI**	**nxCode**
UMI mutados simulados	10 000	10 000	10 000
Corregible de forma única	7703	2447	3829
Dentro de los emparejamientos más cercanos	9242	9779	9629
Tamaño promedio del conjunto más cercano	1,2138	3,0261	2,0978
Dentro de los emparejamientos más cercanos o los segundos más cercanos	9927	9865	9897
Tamaño promedio del segundo conjunto más cercano	3,9391	7,781	6,0504

El conjunto de vNRUMI tiene 120 UMI, de los cuales 50 UMI tienen una longitud de 6 nt y 70 UMI tienen una longitud de 7 nt. El conjunto de NRUMI tiene 218 secuencias de longitud 6. Un planteamiento convencional nxCode utiliza un conjunto de NRUMI de 114 secuencias de longitud de 6 nt. El tamaño promedio de un conjunto es el número promedio de secuencias únicas incluidas en un conjunto.

En la Tabla 2, una corrección única se define como un caso en el que el conjunto de vecinos más cercanos solo tiene una secuencia en él; en otras palabras, el algoritmo de emparejamiento y corrección de UMI descrito anteriormente dio una sugerencia inequívoca del verdadero vNRUMI más probable. Tenga en cuenta que el número de tales secuencias corregibles de forma única es mucho mayor para la metodología de vNRUMI que para NRUMI y nxCode. También, el tamaño promedio del conjunto más cercano/segundo más cercano es mucho menor en el planteamiento de vNRUMI que en otras soluciones, mientras que la tasa en la que el código de barras original no mutado está contenido dentro de esos conjuntos es aproximadamente igual. Esto es importante porque durante la agrupación de lecturas, la información contextual se utiliza para seleccionar un UMI correcto de estos conjuntos más cercano/segundo más cercano. Si esta etapa de agrupación de lecturas con menos secuencias incorrectas puede disminuir la posibilidad de que se haga una elección incorrecta, mejorando en última instancia la capacidad de suprimir el ruido y detectar variantes.

Vale la pena señalar que los planteamientos de NRUMI y nxCode, al igual que otras estrategias de códigos de barras anteriores, presuponen que las secuencias de códigos de barras son todas de longitud uniforme. En la producción de esta simulación, para proporcionar comparaciones directas entre los tres planteamientos, no se utilizaron los métodos originales para corregir errores descritos por los planteamientos de NRUMI y nxCode, lo que podría haber limitado el rendimiento de los planteamientos de NRUMI y nxCode. Sin embargo, los datos de la Tabla 2 proporcionan una idea de la capacidad potencial del planteamiento de vNRUMI para mejorar la corrección de errores, que se ilustra además en el siguiente ejemplo.

Ejemplo 2

Recuperación de fragmentos de ADN mediante vNRUMI y NRUMI

En otro conjunto de ensayos *in silicio*, se prueban las capacidades de vNRUMI y NRUMI para recuperar lecturas. Los ensayos eligen una mutación COSMIC aleatoria y generan un único fragmento de ADN que contiene esa mutación. El tamaño del fragmento tiene un promedio de 166 y una desviación estándar de 40. La simulación agrega un UMI aleatorio a ambos extremos de este fragmento. Utilizó ART (véase, p. ej., https://www.niehs.nih.gov/research/resources/software/biostatistics/art/) para simular 10 lecturas emparejadas por su extremo de esta molécula UMI-fragmento-UMI, y alineó esas lecturas utilizando el alineador Burrows-Wheeler (BWA). Véase, p. ej., http://bio-bwa.sourceforge.net/.

A continuación, el proceso pasa la alineación a un agrupador de lecturas patentado, ReCo, para determinar si puede recuperar la secuencia de fragmentos original y repetir el proceso para lecturas adicionales.

La tabla 3 muestra los números y porcentajes de fragmentos que se pudieron recuperar.

Tabla 3: Tasas de corrección de errores para diseños NRUMI y vNRUMI

Métrica	**Antiguo NRUMI 218**	**Nuevo vNRUMI 120**
Fragmento original perfectamente recuperado	16 837 (95,58 %)	16 915 (96,03 %)
El fragmento original no se recuperó perfectamente	778 (4,42 %)	700 (3,97 %)
Suma	17 615 (100 %)	17 615 (100 %)

El método de vNRUMI recuperó más fragmentos que el método de NRUMI de longitud fija. Una prueba de Chi-square muestra que las diferencias son significativas. χ^2= 4,297, valor P bicola= 0,0382. Utilizando α = 0,05, el método de vNRUMI logró un rendimiento de corrección de errores estadísticamente mejor en comparación con el método de NRUMI, al tiempo que abordaba las deficiencias del método de NRUMI.

La estrategia de NRUMI maneja conjuntos de NRUMI de longitud heterogénea. Esto soluciona el problema de la diversidad de pares de bases que provocó una caída en la calidad de la alineación.

Se proporcionan procesos novedosos para generar conjuntos de UMI de longitud variable que satisfagan las restricciones bioquímicas y para mapear los UMI mal leídos a fin de corregir los UMI. El novedoso planteamiento aborda el problema de la disminución de la calidad de la secuenciación causada por los códigos de barras de longitud uniforme. El uso de un esquema de emparejamiento que tenga en cuenta el número de emparejamientos y emparejamientos erróneos, en lugar de limitarse a rastrear los emparejamientos erróneos, permite mejorar la capacidad de corrección de errores. Las implementaciones son comparables o superan las soluciones existentes, a la vez que proporcionan una funcionalidad adicional.

La presente exposición puede realizarse en otras formas específicas sin apartarse de sus características esenciales. Las realizaciones descritas deben considerarse en todos los aspectos solo como ilustrativas y no restrictivas. El alcance de la invención se indica, por lo tanto, en las reivindicaciones adjuntas en lugar de en la descripción anterior.

REIVINDICACIONES

1. Un método para la secuenciación de moléculas de ácido nucleico de una muestra, que comprende

(a) aplicar adaptadores a los fragmentos de ADN de la muestra para obtener productos adaptadores de ADN,

en donde cada adaptador comprende un índice molecular único no aleatorio,
en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI);

(b) amplificar los productos de ADN-adaptador para obtener una pluralidad de polinucleótidos amplificados;
(c) secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI;
(d) identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo vNRUMI; y
(e) determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI,

en donde la identificación de las lecturas asociadas con el mismo vNRUMI comprende obtener, para cada lectura de la pluralidad de lecturas, puntuaciones de alineación con respecto al conjunto de vNRUMI, indicando cada puntuación de alineación la similitud entre una subsecuencia de una lectura y un vNRUMI, en donde la subsecuencia está en una región de la lectura en la que probablemente se ubiquen los nucleótidos derivados del vNRUMI, y
en donde las puntuaciones de alineación se basan en los emparejamientos de nucleótidos y en las modificaciones de nucleótidos entre la subsecuencia de la lectura y el vNRUMI.

2. El método de la reivindicación 1, en donde cada vNRUMI del conjunto difiere de todos los demás vNRUMI del conjunto en al menos una distancia de modificación definida.

3. El método de la reivindicación 2, en donde la distancia de modificación definida es 3.

4. El método de la reivindicación 1, en donde las modificaciones de nucleótidos comprenden sustituciones, adiciones y deleciones de nucleótidos.

5. El método de la reivindicación 1, en donde cada puntuación de alineación penaliza los emparejamientos erróneos al principio de una subsecuencia, pero no penaliza los emparejamientos erróneos al final de la subsecuencia; opcionalmente

en donde la obtención de una puntuación de alineación entre una lectura y un vNRUMI comprende:

(a) calcular una puntuación de alineación entre el vNRUMI y cada una de las posibles secuencias de prefijo de la subsecuencia de la lectura;
(b) calcular una puntuación de alineación entre la subsecuencia de la lectura y cada una de las posibles secuencias de prefijo del vNRUMI; y
(c) obtener una puntuación de alineación más alta entre las puntuaciones de alineación calculadas en (a) y (b) como la puntuación de alineación entre la lectura y el vNRUMI.

6. El método de la reivindicación 1, en donde la subsecuencia tiene una longitud que es igual a la longitud del vNRUMI más largo del conjunto de vNRUMI.

7. El método de la reivindicación 1, en donde la identificación de las lecturas asociadas al mismo vNRUMI en (d) comprende, además:

seleccionar, para cada lectura de la pluralidad de lecturas, al menos un vNRUMI del conjunto de vNRUMI basándose en las puntuaciones de alineación; y
asociar cada lectura de la pluralidad de lecturas con al menos el vNRUMI seleccionado para la lectura.

8. El método de la reivindicación 7, en donde la selección de al menos un vNRUMI del conjunto de vNRUMI comprende seleccionar un vNRUMI que tenga una puntuación de alineación más alta de entre el conjunto de vNRUMI.

9. El método de la reivindicación 7, en donde al menos un vNRUMI comprende dos o más vNRUMI; opcionalmente comprende además la selección de uno de los dos o más vNRUMI como el mismo vNRUMI de (d) y (e).

10. El método de la reivindicación 1, en donde los adaptadores aplicados en (a) se obtienen mediante:

(i) proporcionar un conjunto de secuencias de oligonucleótidos que tienen al menos dos longitudes moleculares diferentes;
(ii) seleccionar un subconjunto de secuencias de oligonucleótidos del conjunto de secuencias de oligonucleótidos, todas las distancias de modificación entre las secuencias de oligonucleótidos del subconjunto de secuencias de oligonucleótidos que cumplen un valor umbral, el subconjunto de secuencias de oligonucleótidos que forman el conjunto de vNRUMI; y
(iii) sintetizar los adaptadores, cada uno de los cuales comprende una región hibridada bicatenaria, un brazo 5' monocatenario, un brazo 3' monocatenario y al menos un vNRUMI del conjunto de vNRUMI; opcionalmente

en donde el valor umbral es 3.

11. El método de la reivindicación 1, en donde el conjunto de vNRUMI comprende vNRUMI de 6 nucleótidos y vNRUMI de 7 nucleótidos.

12. El método de la reivindicación 1, en donde (e) comprende:

(1) agrupar las lecturas asociadas con el mismo vNRUMI en un grupo para obtener una secuencia de nucleótidos de consenso para la secuencia del fragmento de ADN de la muestra; opcionalmente, en donde la secuencia de nucleótidos de consenso se obtiene basándose parcialmente en las puntuaciones de calidad de las lecturas; o
(2) identificar, entre las lecturas asociadas con el mismo vNRUMI, las lecturas que tienen una misma posición de lectura o posiciones de lectura similares en una secuencia de referencia, y determinar la secuencia del fragmento de ADN usando lecturas que (i) están asociadas con el mismo vNRUMI y (ii) tienen la misma posición de lectura o posiciones de lectura similares en la secuencia de referencia.

13. El método de la reivindicación 1, en donde el conjunto de vNRUMI incluye no más de aproximadamente 10 000, aproximadamente 1000 o aproximadamente 200 vNRUMI diferentes.

14. El método de la reivindicación 1, donde la aplicación de adaptadores a los fragmentos de ADN de la muestra comprende la aplicación de adaptadores a ambos extremos de los fragmentos de ADN de la muestra.

15. Un sistema informático, que comprende:

uno o más procesadores;
memoria de sistema; y
uno o más soportes de almacenamiento legibles por ordenador en los que se han almacenado sobre los mismos instrucciones ejecutables por ordenador que hacen que el sistema informático implemente un método para determinar la información de secuencia de una secuencia de interés de una muestra, comprendiendo las instrucciones:

(a) obtener una pluralidad de lecturas de una pluralidad de polinucleótidos amplificados, comprendiendo cada polinucleótido de la pluralidad de polinucleótidos amplificados un adaptador unido a un fragmento de ADN,

en donde el adaptador comprende un índice molecular único no aleatorio, y
en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes, formando un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI);

(b) identificar, entre la pluralidad de lecturas, las lecturas asociadas con un mismo vNRUMI, en donde la identificación de las lecturas asociadas con el mismo vNRUMI comprende la obtención, para cada lectura de la pluralidad de lecturas, de puntuaciones de alineación con respecto al conjunto de vNRUMI, cada puntuación de alineación indica la similitud entre una subsecuencia de una lectura y un vNRUMI, en donde la subsecuencia se encuentra en una región de la lectura en la que es probable que se ubiquen los nucleótidos derivados del vNRUMI, en donde las puntuaciones de alineación se basan en emparejamientos de nucleótidos y modificaciones de nucleótidos entre la subsecuencia de la lectura y el vNRUMI; y

(c)determinar, mediante las lecturas asociadas con el mismo vNRUMI, una secuencia de un fragmento de ADN de la muestra.

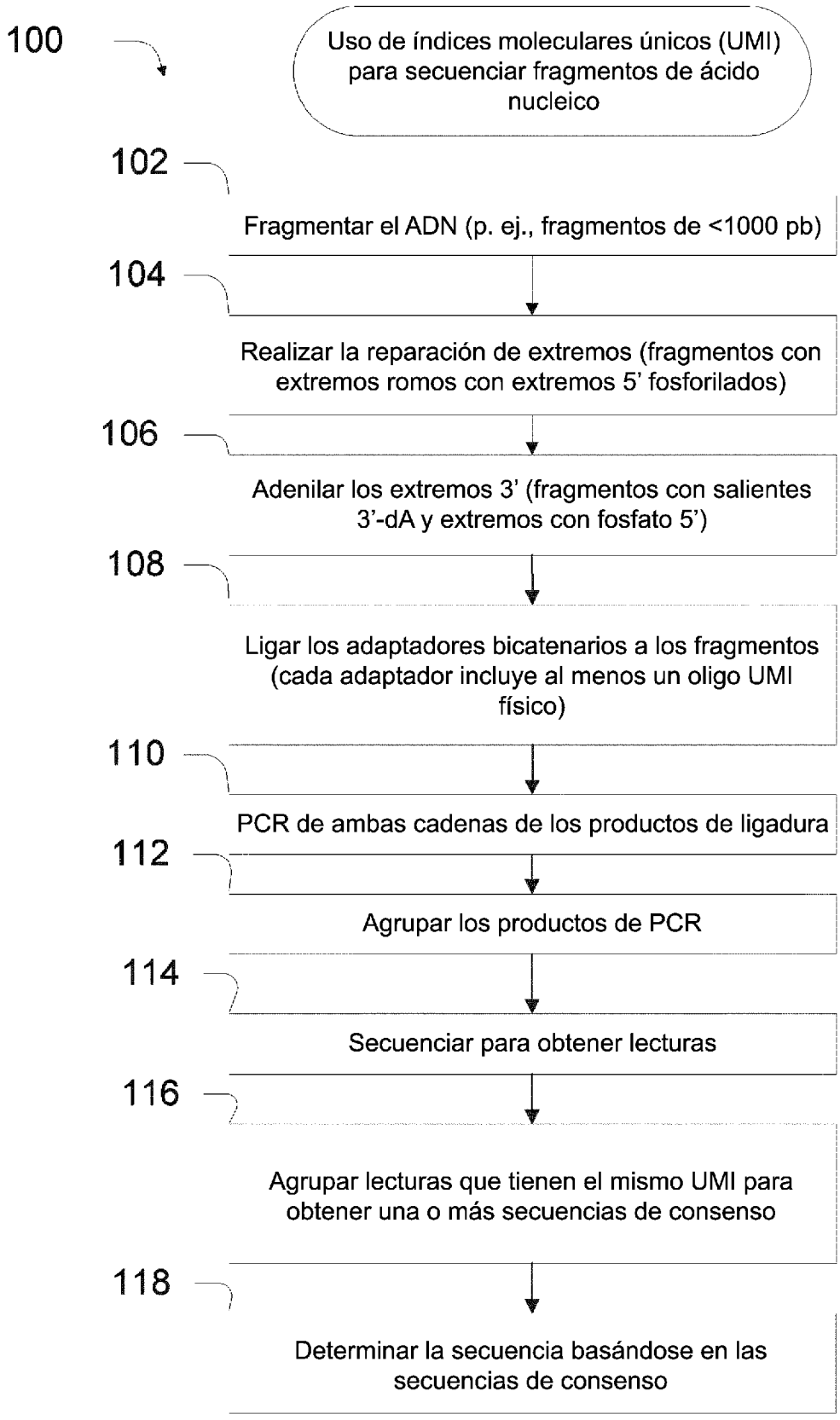

100
Uso de índices moleculares únicos (UMI) para secuenciar fragmentos de ácido nucleico
102
Fragmentar el ADN (p. ej., fragmentos de <1000 pb)
104
Realizar la reparación de extremos (fragmentos con extremos romos con extremos 5' fosforilados)
106
Adenilar los extremos 3' (fragmentos con salientes 3'-dA y extremos con fosfato 5')
108
Ligar los adaptadores bicatenarios a los fragmentos (cada adaptador incluye al menos un oligo UMI físico)
110
PCR de ambas cadenas de los productos de ligadura
112
Agrupar los productos de PCR
114
Secuenciar para obtener lecturas
116
Agrupar lecturas que tienen el mismo UMI para obtener una o más secuencias de consenso
118
Determinar la secuencia basándose en las secuencias de consenso

Figura 1A

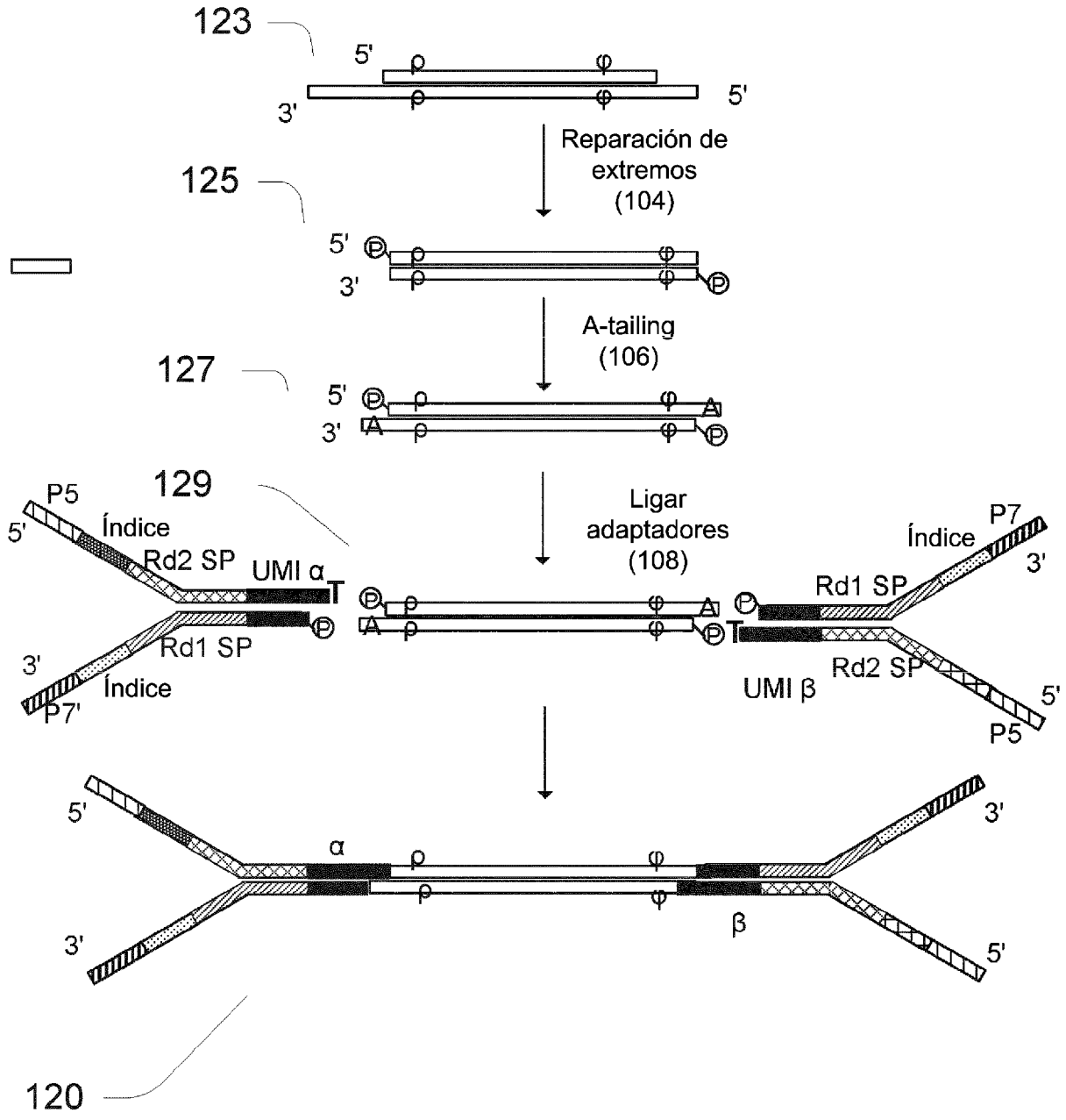

123
5'
3'
5'
Reparación de extremos (104)
125
5'
3'
A-tailing (106)
127
5'
3'
129
Ligar adaptadores (108)
P5
5'
Índice
Rd2 SP
UMI α
Rd1 SP
3'
Índice
P7'
P7
Índice
3'
Rd1 SP
Rd2 SP
UMI β
5'
P5
5'
3'
α
β
3'
5'
120

Figura 1B

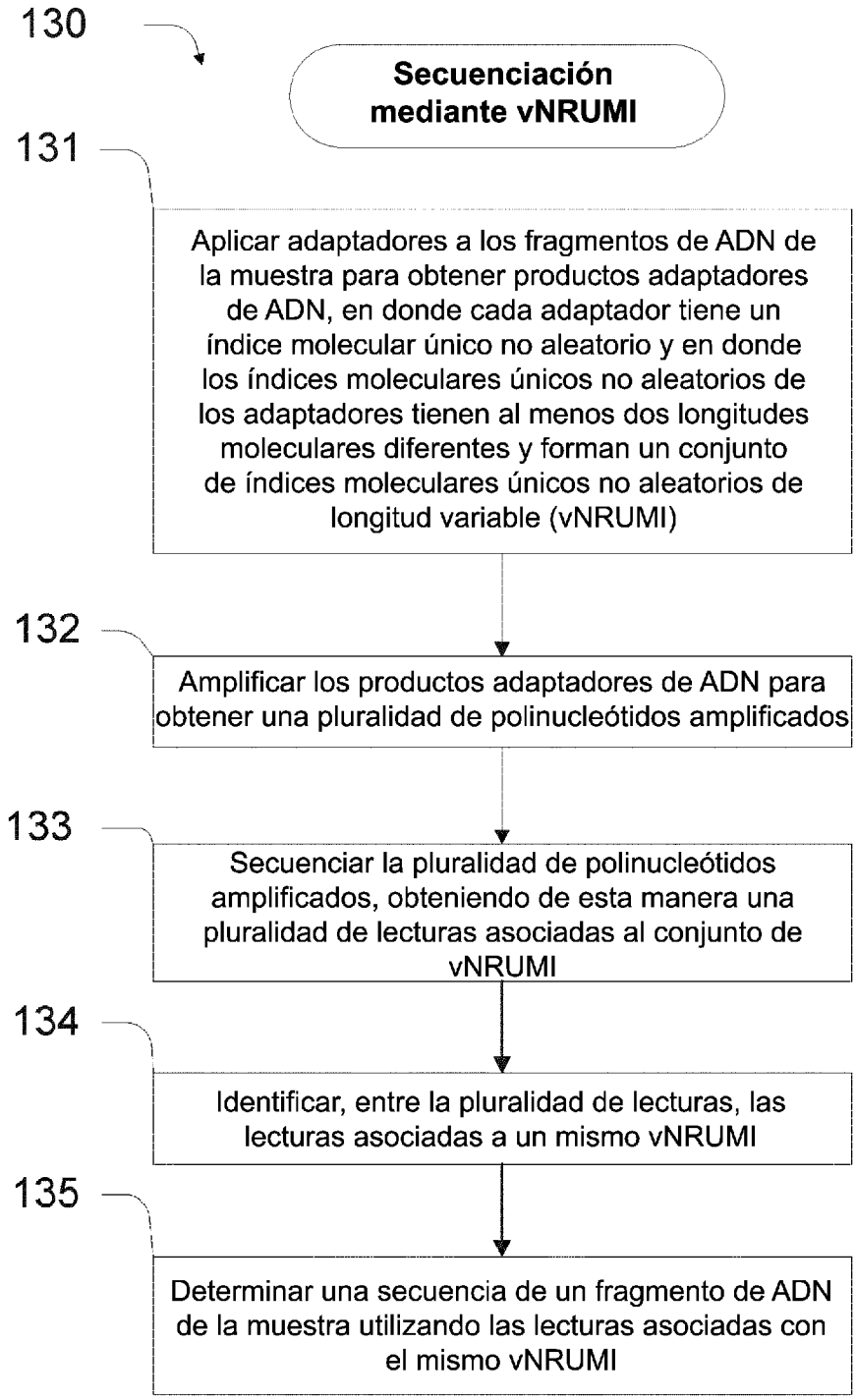

130
Secuenciación mediante vNRUMI
131
Aplicar adaptadores a los fragmentos de ADN de la muestra para obtener productos adaptadores de ADN, en donde cada adaptador tiene un índice molecular único no aleatorio y en donde los índices moleculares únicos no aleatorios de los adaptadores tienen al menos dos longitudes moleculares diferentes y forman un conjunto de índices moleculares únicos no aleatorios de longitud variable (vNRUMI)
132
Amplificar los productos adaptadores de ADN para obtener una pluralidad de polinucleótidos amplificados
133
Secuenciar la pluralidad de polinucleótidos amplificados, obteniendo de esta manera una pluralidad de lecturas asociadas al conjunto de vNRUMI
134
Identificar, entre la pluralidad de lecturas, las lecturas asociadas a un mismo vNRUMI
135
Determinar una secuencia de un fragmento de ADN de la muestra utilizando las lecturas asociadas con el mismo vNRUMI

Figura 1C

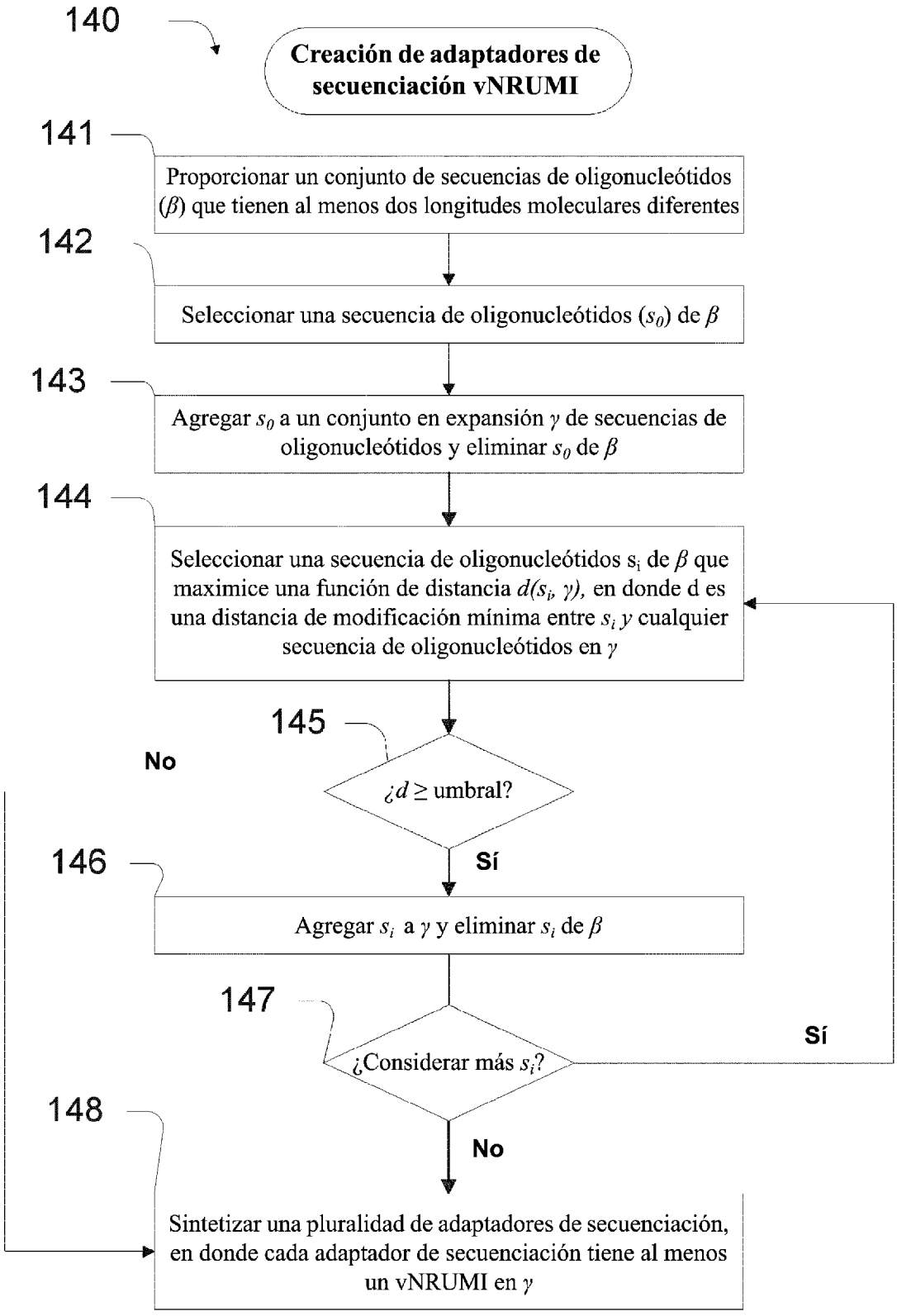

140
Creación de adaptadores de secuenciación vNRUMI
141
Proporcionar un conjunto de secuencias de oligonucleótidos (β) que tienen al menos dos longitudes moleculares diferentes
142
Seleccionar una secuencia de oligonucleótidos (s_0) de β
143
Agregar s_0 a un conjunto en expansión γ de secuencias de oligonucleótidos y eliminar s_0 de β
144
Seleccionar una secuencia de oligonucleótidos s_i de β que maximice una función de distancia $d(s_i, \gamma)$, en donde d es una distancia de modificación mínima entre s_i y cualquier secuencia de oligonucleótidos en γ
145
¿$d \geq$ umbral?
No
Sí
146
Agregar s_i a γ y eliminar s_i de β
147
¿Considerar más s_i?
Sí
No
148
Sintetizar una pluralidad de adaptadores de secuenciación, en donde cada adaptador de secuenciación tiene al menos un vNRUMI en γ

Figura 1D

150

	Q	G	T	C	T	T	C	G
S1	0	-1	-2	-3	-4	-5	-6	**-7**
A	-1	-1	-2	-3	-4	-5	-6	**-7**
A	-2	-2	-2	-3	-4	-5	-6	**-7**
C	-3	-3	-3	-1	-2	-3	-4	**-5**
T	-4	-4	-2	-2	0	-1	-2	**-3**
T	-5	-5	-3	-3	-1	1	0	**-1**
C	**-6**	**-6**	**-4**	**-2**	**-2**	**0**	**2**	**1**

151

```
Q:  GTCTTC_
S1: AACTTC
```
Distancia de Levenshtein (n.º adición + n.º deleción + n.º sustitución) = 2

```
Q:  GTCTTC_
S1: AACTTC
```
Puntuación de alineación glocal (n.º emparejamiento - Distancia de Levenshtein) = 2

Distancia de Levenshtein = 2

152

	Q	G	T	C	T	T	C	G
S2	0	-1	-2	-3	-4	-5	-6	**-7**
C	-1	-1	-2	-3	-4	-5	-6	**-7**
G	-2	0	-1	-2	-2	-3	-4	**-3**
C	-3	-1	-1	0	-1	-2	-2	**-3**
T	-4	-2	0	-1	1	0	-1	**-2**
T	-5	-3	-1	-1	0	2	1	**0**
C	-6	-4	-2	0	-1	1	3	**2**
G	**-7**	**-5**	**-3**	**-1**	**-1**	**0**	**2**	**4**

153

```
Q:  GTCTTCG
S2: CGCTTCG
```
Distancia de Levenshtein = 2

```
Q:  GTCTTCG
S2: CGCTTCG
```
Puntuación de alineación glocal = 4

Figura 1E

160

	Q	T	T	G	G	C	A	T
S_1	0	-1	-2	-3	-4	-5	-6	**-7**
T	-1	1	0	-1	-2	-3	-4	**-5**
T	-2	0	2	1	0	-1	-2	**-3**
G	-3	-1	1	3	2	1	0	**-1**
T	-4	-2	0	2	2	1	0	**1**
G	-5	-3	-1	1	3	2	1	**0**
A	-6	-4	-2	0	2	2	3	**2**
C	**-7**	**-5**	**-3**	**-1**	**1**	**3**	**2**	**2**

Q: TTGGCAT
S1: TTGTGAC

Puntuación de alineación glocal = 2
Puntuación de alineación glocal = 3

162 161

163

	Q	T	T	G	G	C	A	T
S_2	0	-1	-2	-3	-4	-5	-6	**-7**
G	-1	-1	-2	-1	-2	-3	-4	**-5**
G	-2	-2	-2	-1	0	-1	-2	**-3**
C	-3	-3	-3	-2	-1	1	0	**-1**
C	-4	-4	-4	-3	-2	0	0	**-1**
A	-5	-5	-5	-4	-3	-1	1	**0**
T	**-6**	**-4**	**-4**	**-5**	**-4**	**-2**	**0**	**2**

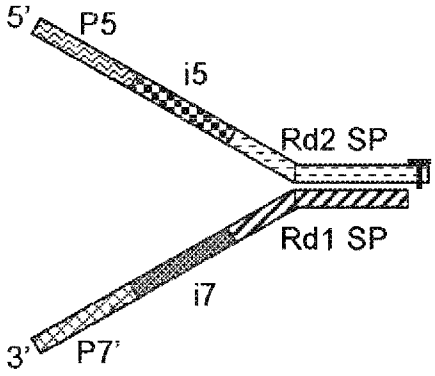
164

Q: TTGGCAT
S2: GGCCAT

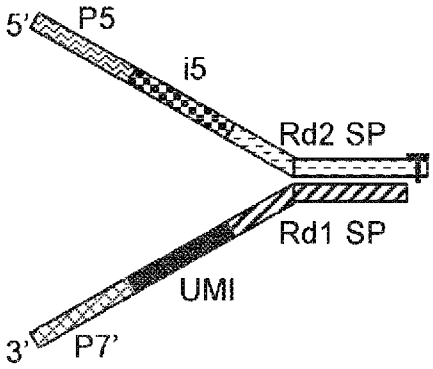
Puntuación de alineación glocal = 2
Puntuación de alineación glocal = 2

Figura 1F

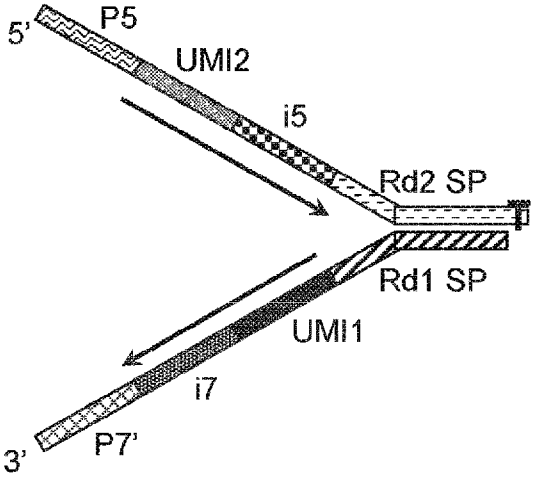
(i) Adaptador de doble índice estándar TruSeq



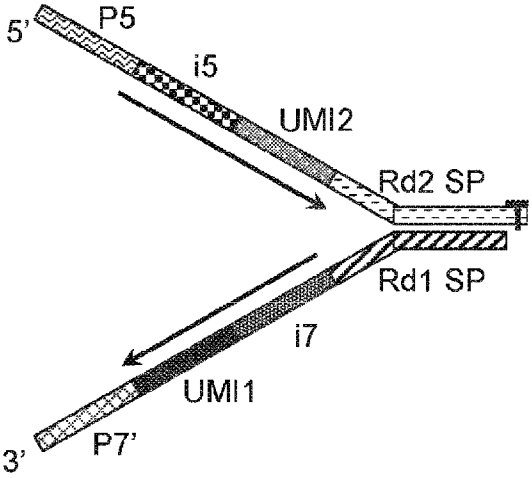
(ii) UMI que reemplaza la posición del índice de muestra 1



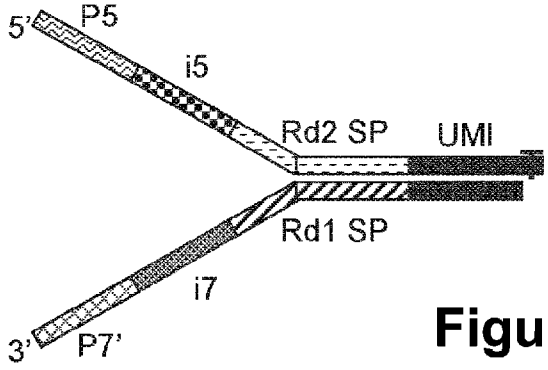
(iii) UMI en los brazos P5 y P7, antes de la lectura de índice de muestra



(iv) UMI en los brazos P5 y P7, detrás de la lectura de índice de muestra



(v) Adaptador de doble índice con un UMI en la región bicatenaria



(vi) Adaptador corto con un UMI en la región bicatenaria

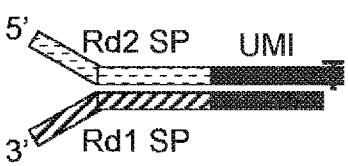


Figura 2A

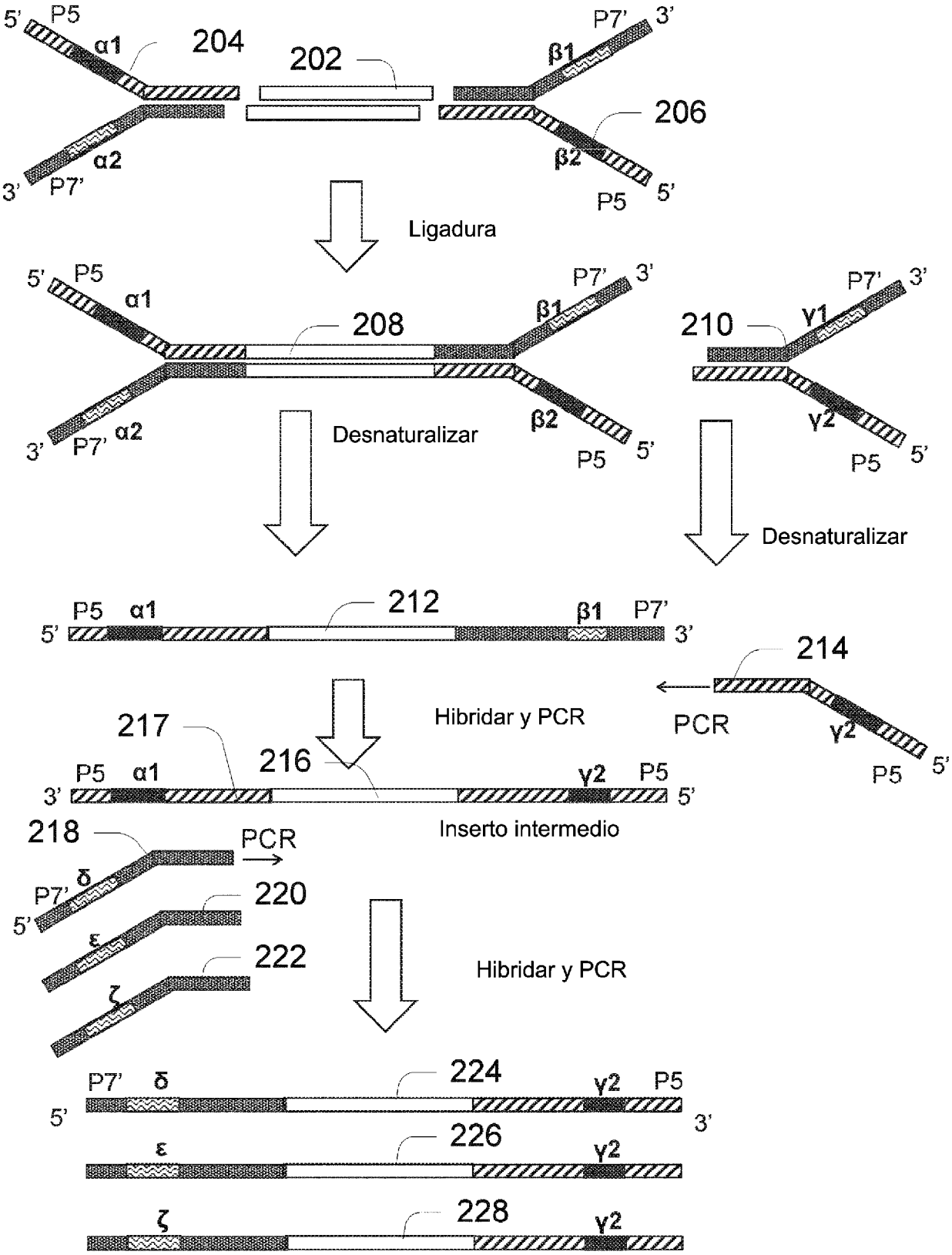

5' P5 α1 204 202 β1 P7' 3'
206
α2 3' P7'
β2 P5 5'
Ligadura
208
210 γ1
Desnaturalizar
Desnaturalizar
212
214
Hibridar y PCR
PCR
217
216
Inserto intermedio
218 PCR
δ
220
ε
222
ζ
Hibridar y PCR
224
226
228
γ2

Figura 2B

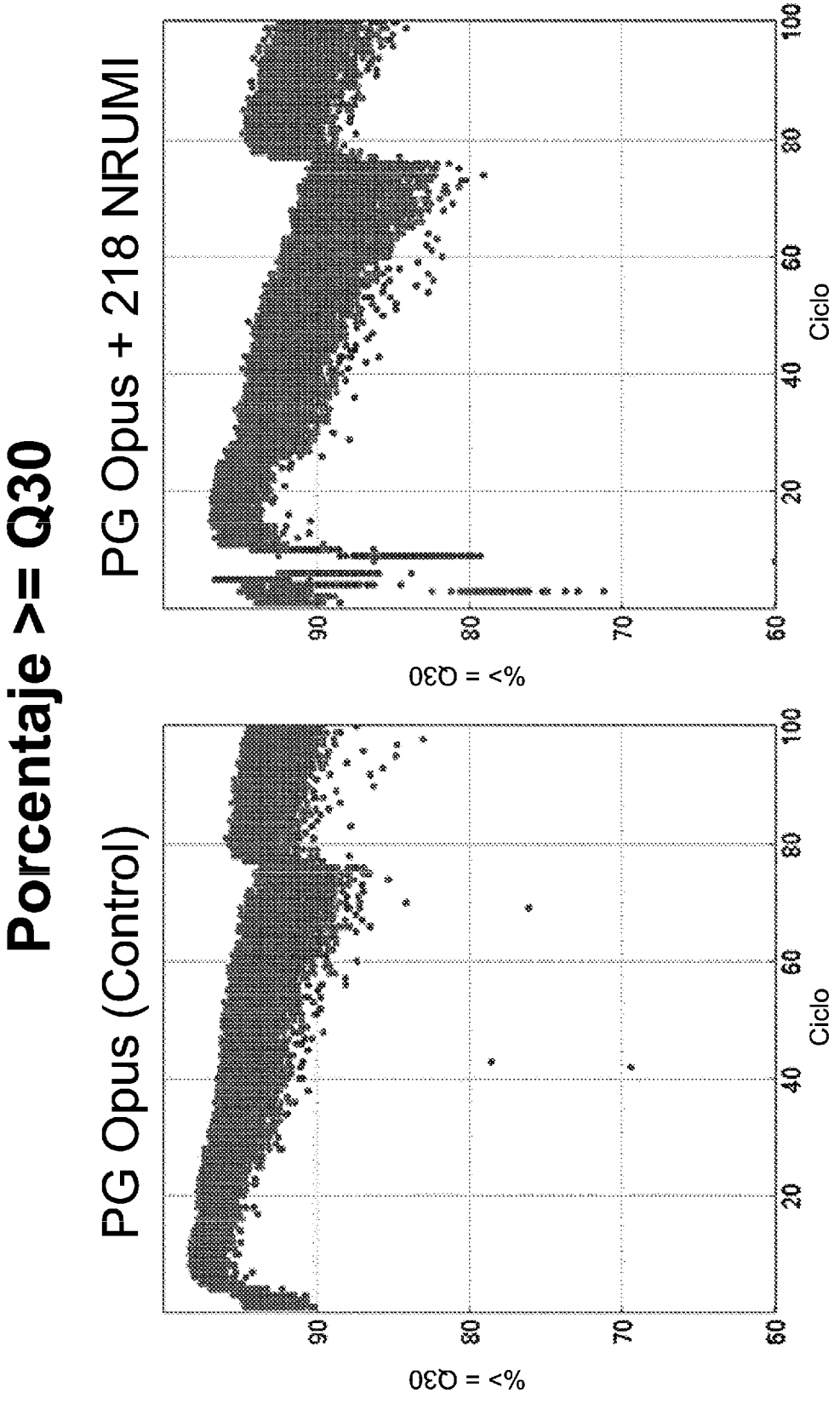

Porcentaje >= Q30
PG Opus (Control)
PG Opus + 218 NRUMI
% >= Q30
Ciclo
60
70
80
90
100
20
40
60
80

Figura 2C

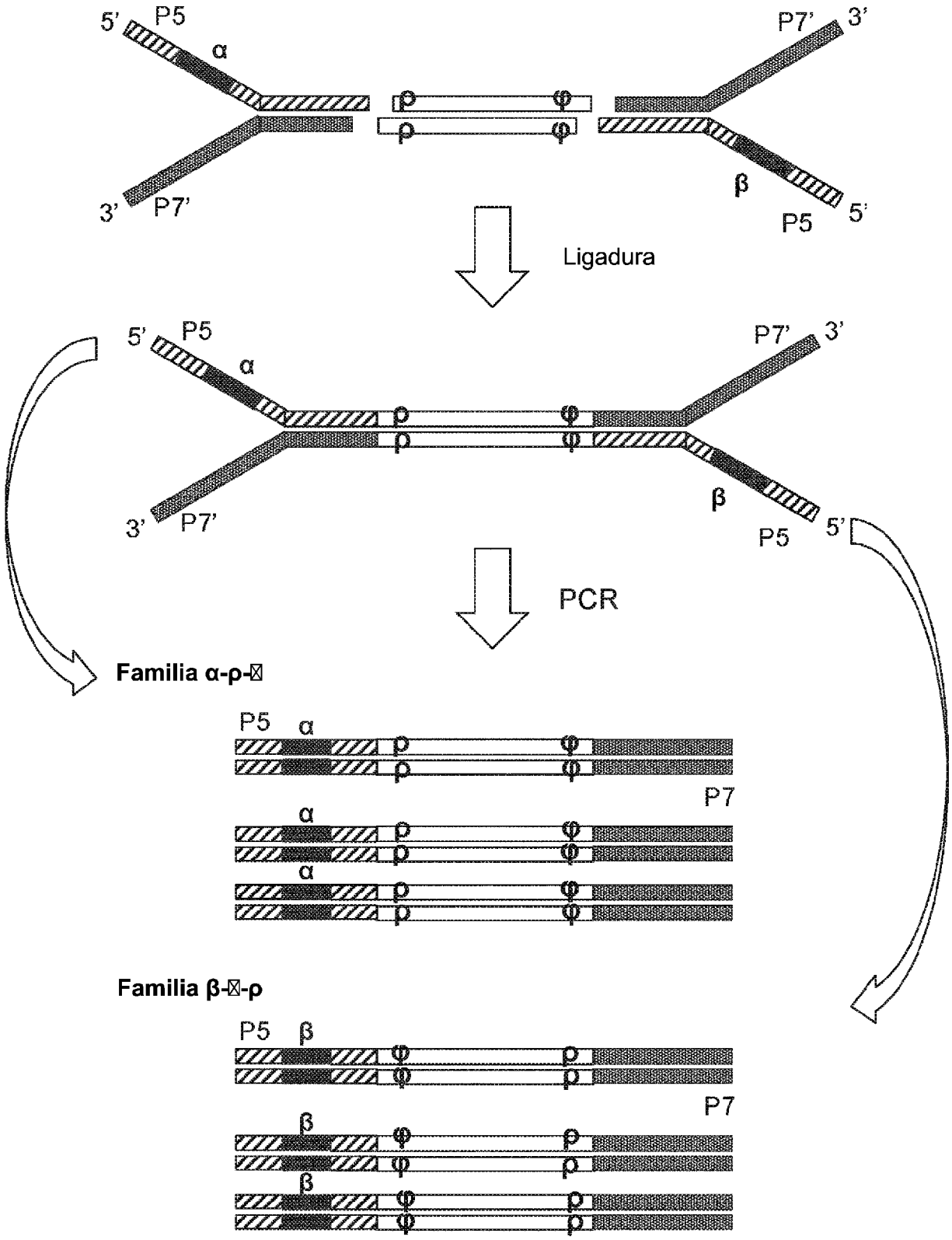

5'
P5
α
P7'
3'
3'
P7'
β
P5
5'
Ligadura
5'
P5
α
P7'
3'
3'
P7'
β
P5
5'
PCR
Familia α-ρ-φ
P5
α
P7
α
α
Familia β-φ-ρ
P5
β
P7
β
β

Figura 3A

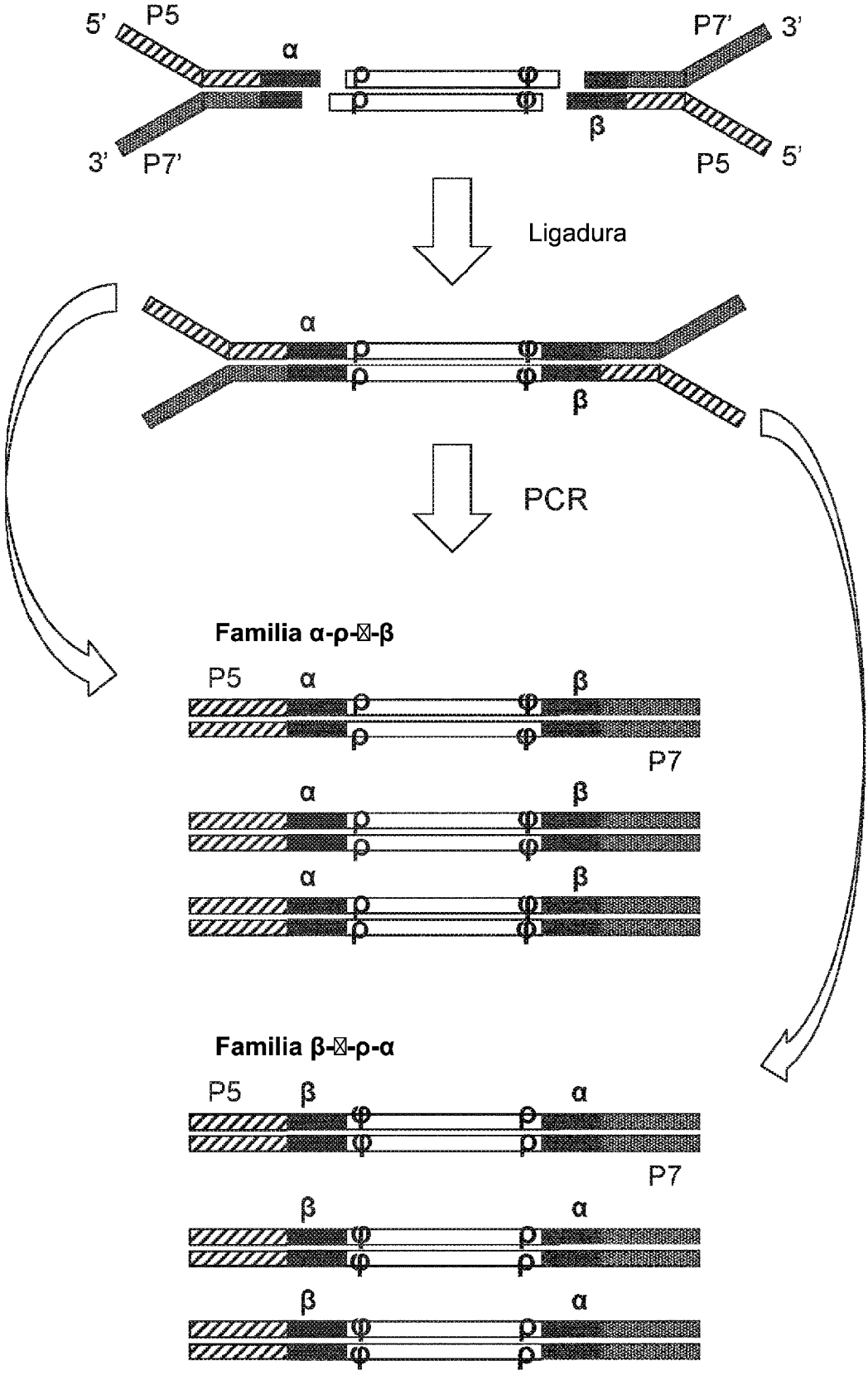

5'
P5
α
P7'
3'
3'
P7'
β
P5
5'
Ligadura
α
β
PCR
Familia α-ρ-φ-β
P5
α
β
P7
α
β
α
β
Familia β-φ-ρ-α
P5
β
α
P7
β
α
β
α

Figura 3B

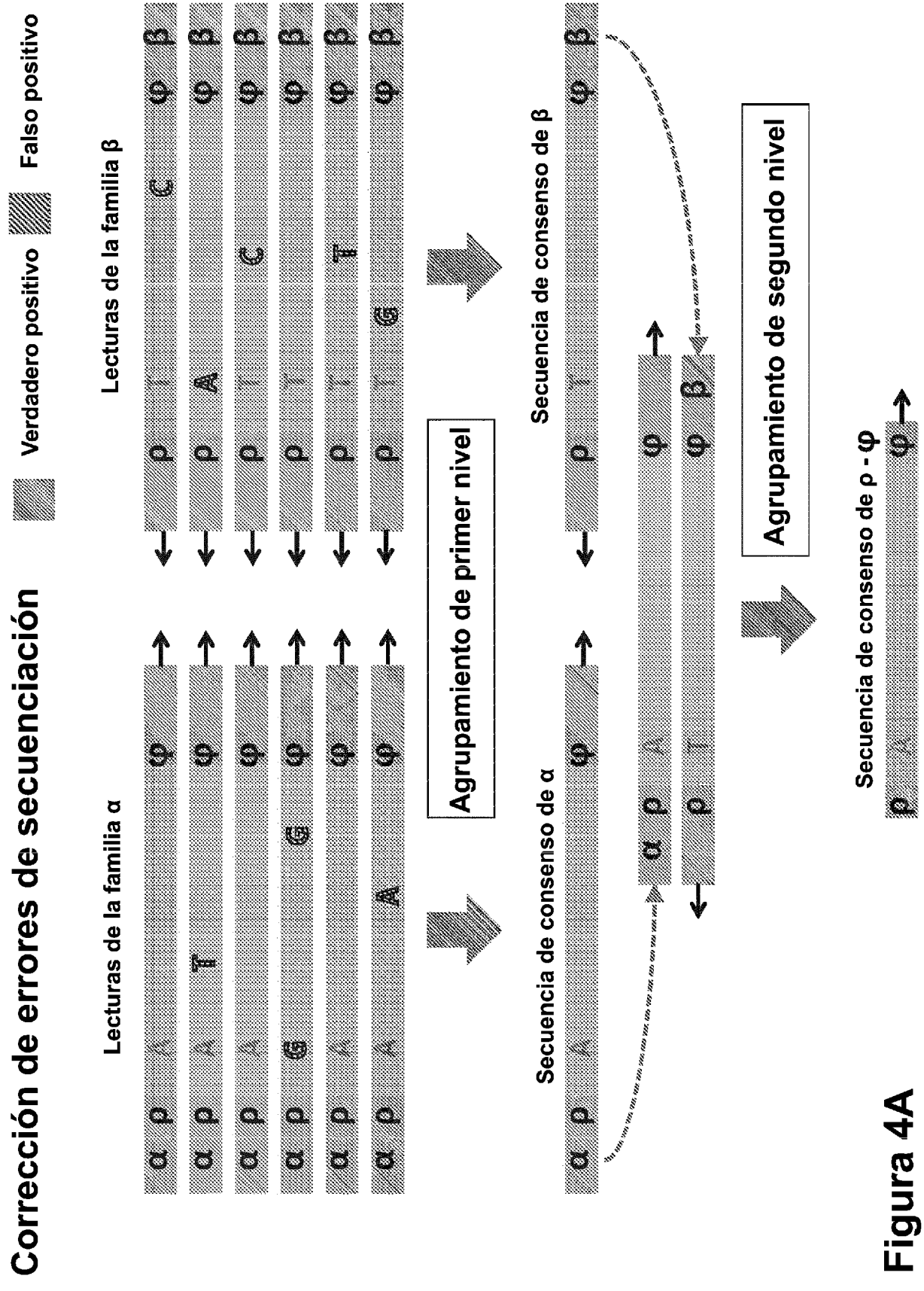

Corrección de errores de secuenciación
Verdadero positivo
Falso positivo
Lecturas de la familia α
Lecturas de la familia β
Agrupamiento de primer nivel
Secuencia de consenso de α
Secuencia de consenso de β
Agrupamiento de segundo nivel
Secuencia de consenso de ρ - φ

Figura 4A

Corrección de errores de la PCR y la secuenciación

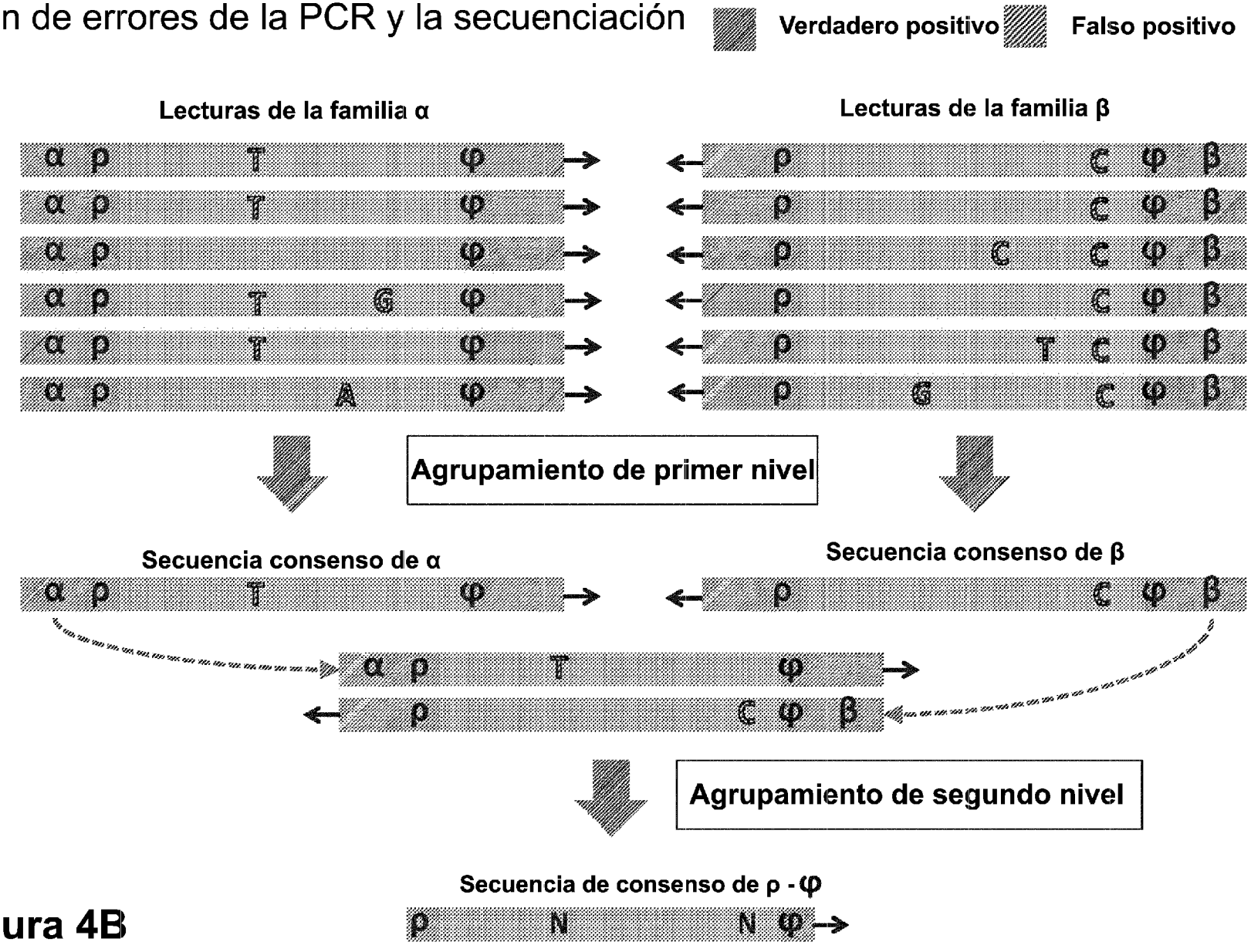


Figura 4B

Corrección de homopolímeros

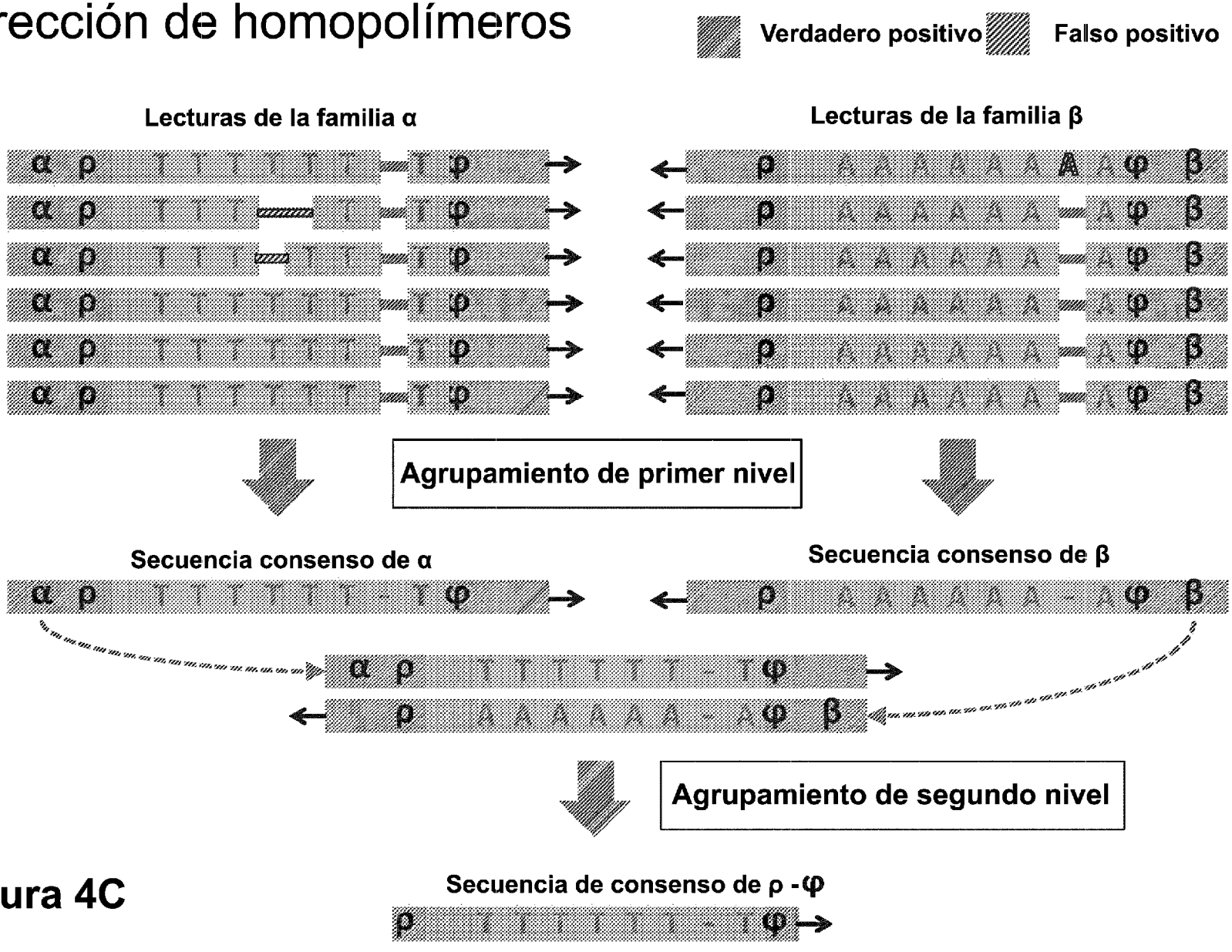


Figura 4C

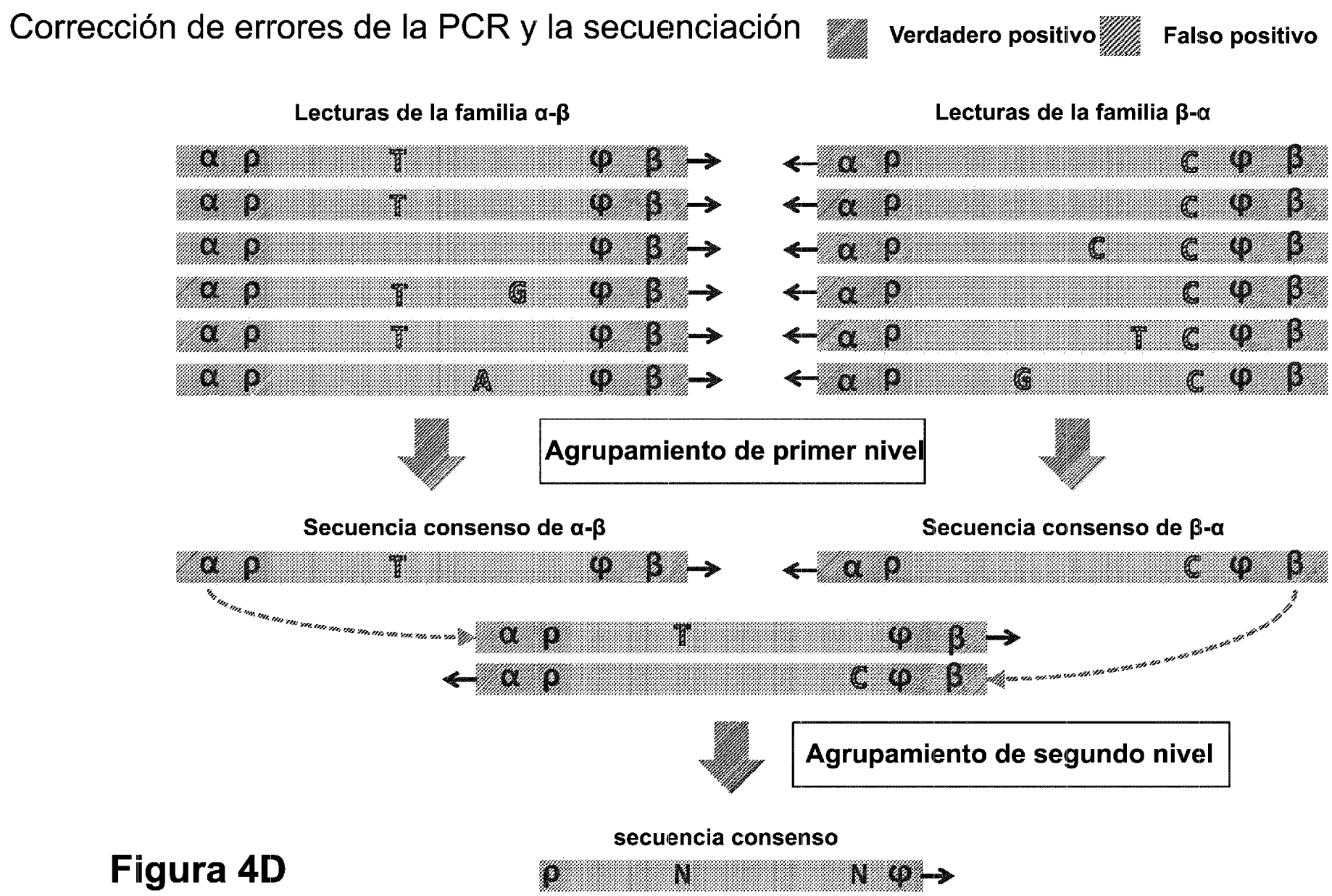

Corrección de errores de la PCR y la secuenciación
Verdadero positivo
Falso positivo
Lecturas de la familia α-β
Lecturas de la familia β-α
Agrupamiento de primer nivel
Secuencia consenso de α-β
Secuencia consenso de β-α
Agrupamiento de segundo nivel
secuencia consenso

Figura 4D

Corrección de homopolímeros

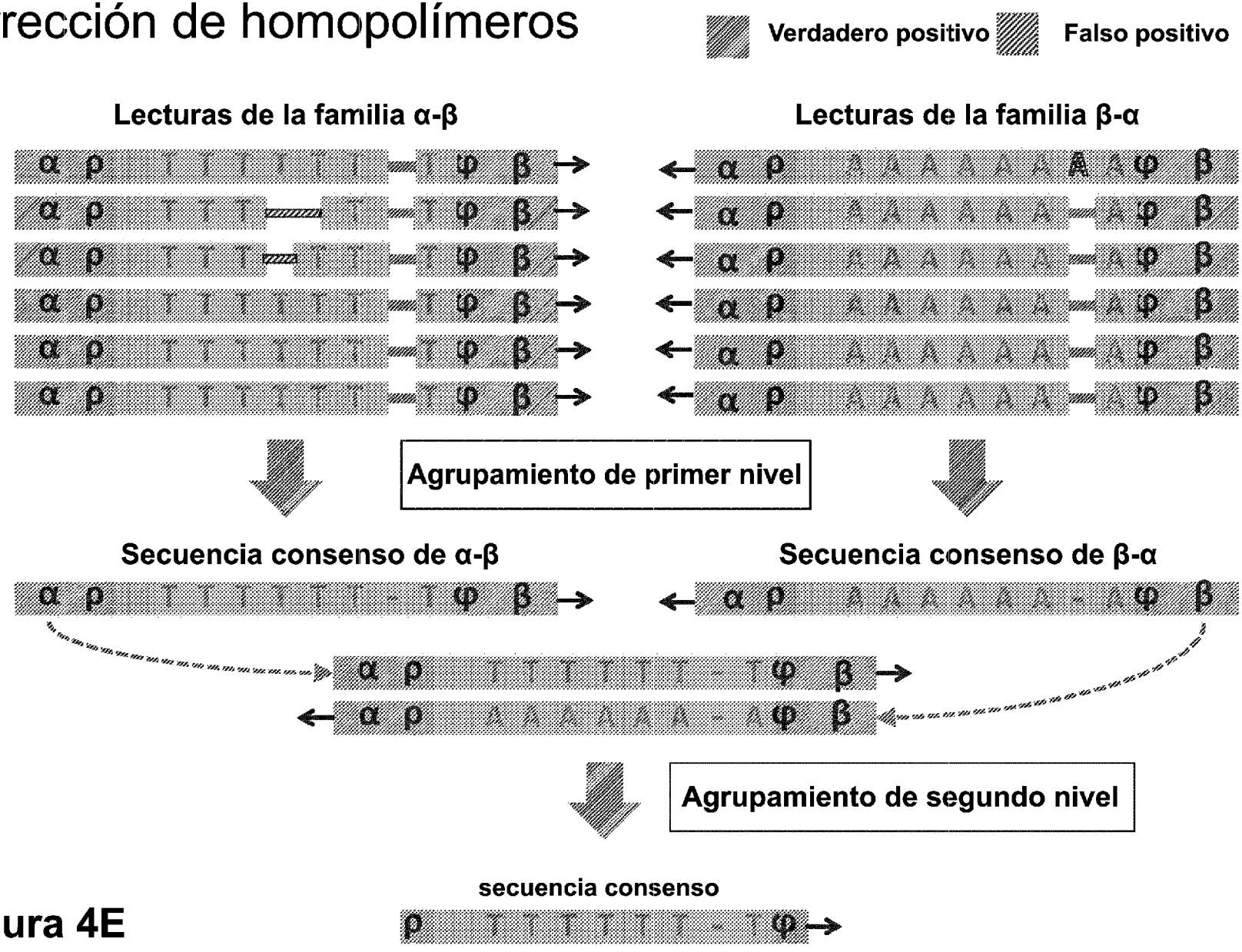


Figura 4E

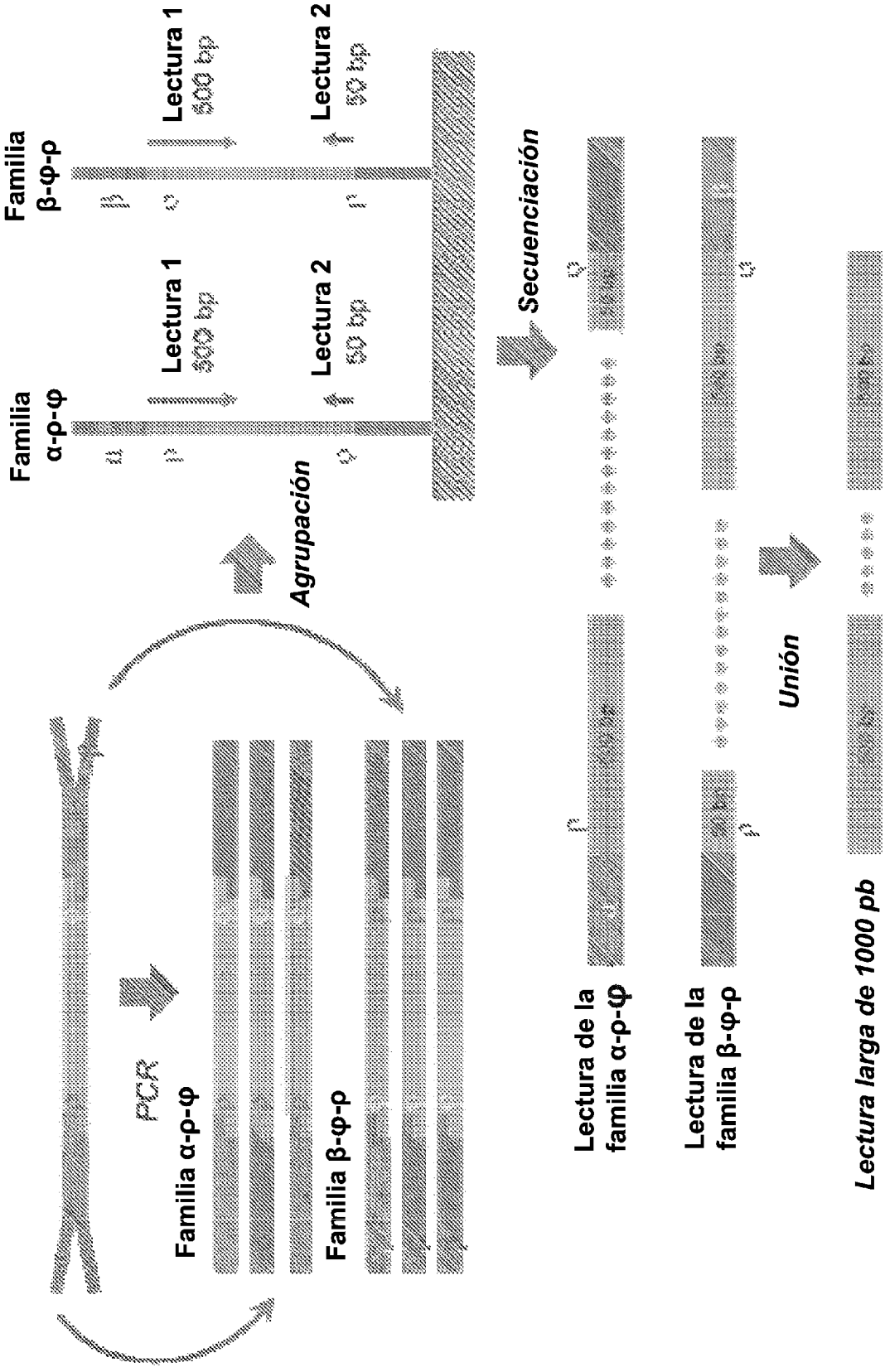

PCR
Familia α-ρ-φ
Familia β-φ-ρ
Agrupación
Familia α-ρ-φ
Lectura 1
600 bp
Lectura 2
50 bp
Familia β-φ-ρ
Lectura 1
600 bp
Lectura 2
50 bp
Secuenciación
Lectura de la familia α-ρ-φ
Lectura de la familia β-φ-ρ
Unión
Lectura larga de 1000 pb

Figura 5

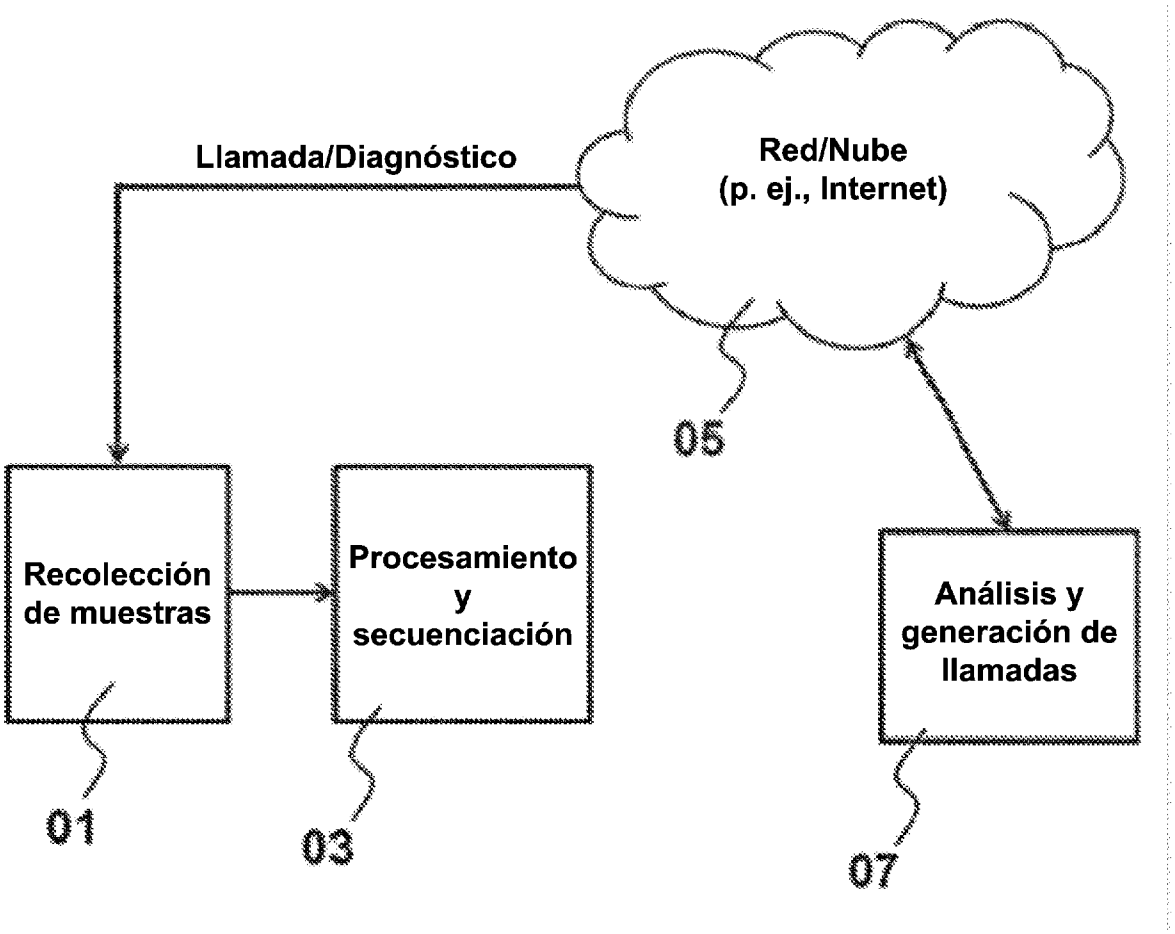

Llamada/Diagnóstico
Red/Nube
(p. ej., Internet)
05
Recolección
de muestras
Procesamiento
y
secuenciación
Análisis y
generación de
llamadas
01
03
07

Figura 6

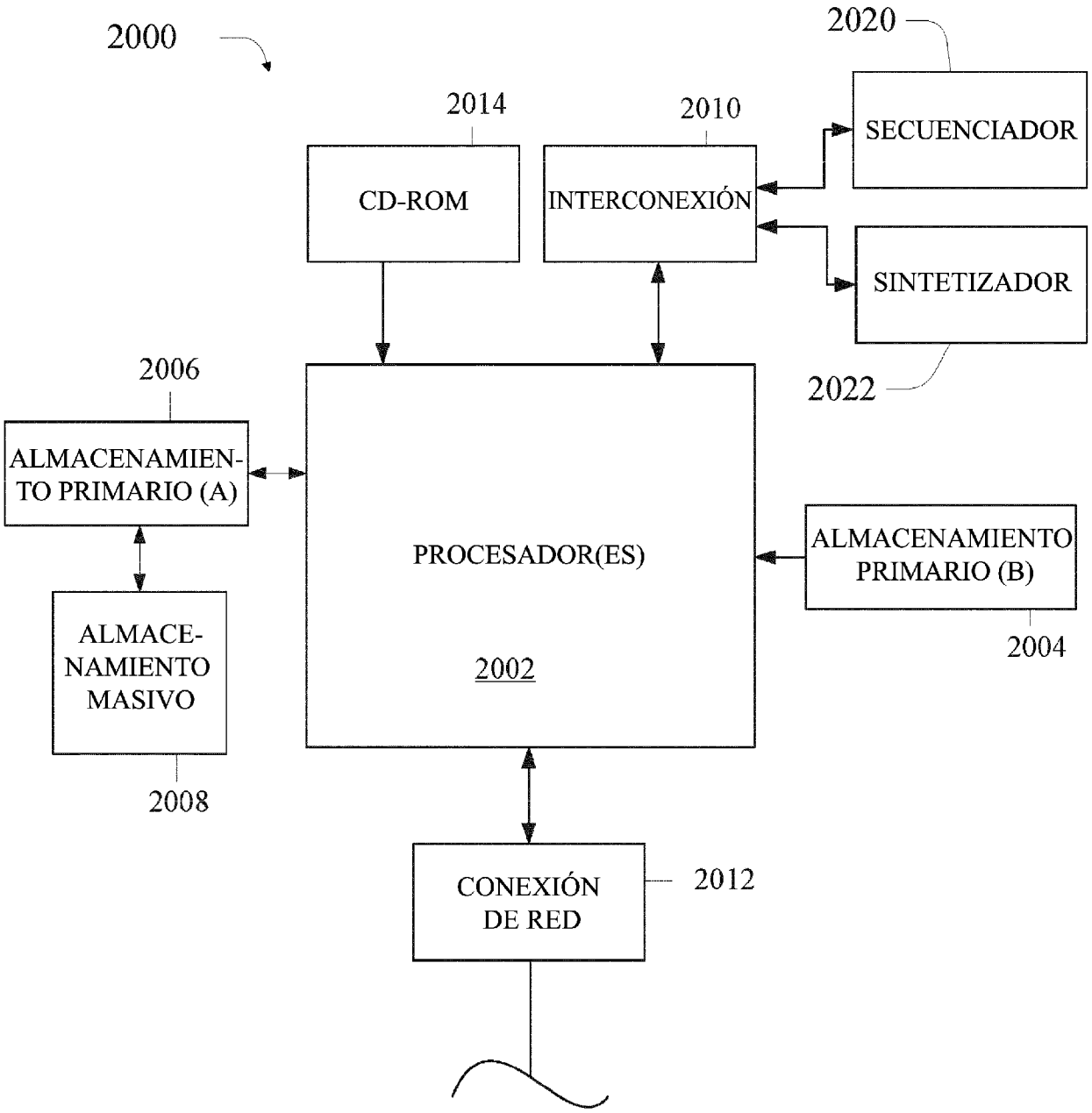

2000
2020
2014
2010
SECUENCIADOR
CD-ROM
INTERCONEXIÓN
SINTETIZADOR
2022
2006
ALMACENAMIEN-
TO PRIMARIO (A)
PROCESADOR(ES)
2002
ALMACENAMIENTO
PRIMARIO (B)
2004
ALMACE-
NAMIENTO
MASIVO
2008
CONEXIÓN
DE RED
2012

Figura 7