

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 966 831**

51 Int. Cl.:

G16B 35/10 (2009.01)

C12Q 1/6806 (2008.01)

C12Q 1/6869 (2008.01)

C40B 40/06 (2006.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **05.11.2018** **PCT/US2018/059255**

87 Fecha y número de publicación internacional: **09.05.2019** **WO19090251**

96 Fecha de presentación y número de la solicitud europea: **05.11.2018** **E 18804892 (0)**

97 Fecha y número de publicación de la concesión europea: **25.10.2023** **EP 3707723**

54 Título: **Técnicas de indexación de ácidos nucleicos**

30 Prioridad:

06.11.2017 US 201762582175 P

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

24.04.2024

73 Titular/es:

ILLUMINA, INC. (100.0%)
5200 Illumina Way
San Diego, CA 92122, US

72 Inventor/es:

VIECELI, JOHN S. y
KELLEY, RYAN MATTHEW

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

ES 2 966 831 T3

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Técnicas de indexación de ácidos nucleicos

5 Referencia cruzada a solicitudes relacionadas

La presente solicitud reivindica la prioridad y el beneficio de la solicitud provisional US 62/582.175, presentada el 6 de noviembre de 2017, titulada "NUCLEIC ACID INDEXING TECHNIQUES" ("TÉCNICAS DE INDEXACIÓN DE ÁCIDOS NUCLEICOS").

10 Antecedentes

La presente descripción se refiere en general al campo de datos relacionados con muestras biológicas, tales como datos de secuencias. Más particularmente, la descripción se refiere a técnicas para indexar ácidos nucleicos y resolver secuencias indexadas en datos de secuencias adquiridos.

La secuenciación genética se ha convertido en un área cada vez más importante de la investigación genética y promete usos futuros en diagnóstico y otras aplicaciones. En general, la secuenciación genética implica determinar el orden de los nucleótidos para un ácido nucleico, tal como un fragmento de ADN o ARN. La tecnología de secuenciación de próxima generación facilita una secuenciación de mayor rendimiento mediante el uso de preparaciones de muestras agrupadas, p. ej., preparaciones de muestras múltiples. Dentro de las preparaciones de muestras agrupadas, cada muestra individual puede etiquetarse o marcarse de otro modo para que cada lectura de secuenciación de la muestra agrupada pueda asociarse o atribuirse a una muestra individual del conjunto. Sin embargo, los datos de secuencias adquiridos pueden tener errores y ruido introducidos desde diversas fuentes, p. ej., defectos de las muestras, errores de preparación de las muestras y sesgo de secuenciación, lo que puede disminuir la precisión de los datos de secuencias asociados a cada muestra individual. Por lo tanto, es deseable desarrollar métodos que permitan una secuenciación de alto rendimiento en una preparación de múltiples muestras con errores de atribución reducidos. Kircher M, y col. (2011) "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform", Nucleic Acids Res., 2012, Vol. 40, n.º 1, e3, describe la indexación doble. Meyer M, y col. (2007), "Targeted high-throughput sequencing of tagged nucleic acid samples", Nucleic Acids Res., 2007, Vol. 35, n.º 15, e97, describe la secuenciación con marcadores en paralelo.

Breve descripción

La presente descripción proporciona un enfoque novedoso para la indexación de muestras de una muestra biológica. Como se proporciona en el presente documento, la indexación de muestras es el proceso mediante el cual las lecturas de secuenciación adquiridas se asocian con una muestra particular. En particular, la técnica de indexación descrita no introduce contaminación entre muestras, funciona de manera sólida con números arbitrariamente bajos de muestras (hasta una para el seguimiento positivo de la muestra) y admite una representación uniforme entre muestras múltiples. Los problemas de contaminación cruzada de índices, rendimiento de indexación de muestras de baja multiplicidad y rendimiento de índice variable se resuelven con una estrategia de indexación que utiliza indexación doble por cada fragmento para marcar cada fragmento de ácido nucleico preparado a partir de una muestra utilizando un posible conjunto de índices múltiples para el primer y segundo índice. La técnica descrita asigna una pluralidad de índices únicos a cada muestra tanto en el primer como en el segundo índice, a diferencia del enfoque combinatorio tradicional. En una realización, una secuencia índice es una secuencia de ácido nucleico exógena que puede incorporarse a ácidos nucleicos de una muestra para la identificación de la muestra y/o asociación de la muestra. Por ejemplo, las secuencias índice proporcionadas en el presente documento pueden incorporarse en fragmentos de ácido nucleico procedentes de muestras de interés durante la preparación de las muestras y/o durante la secuenciación. Cada secuencia índice tiene una secuencia distinguible con respecto a otras secuencias índice utilizadas en la reacción de secuenciación. En determinadas realizaciones, las secuencias índice se pueden distinguir de los ácidos nucleicos endógenos de la muestra o muestras. Las secuencias índice pueden ser monocatenarias o bicatenarias y pueden tener al menos 5 bases, al menos 6 bases, al menos 7 bases, al menos 8 bases o más. En determinadas realizaciones, las secuencias índice tienen una longitud de 5-8 bases, 5-10 bases, 5-15 bases, 5-25 bases, 8-10 bases, 8-12 bases, 8-15 bases u 8-25 bases, etc.

Las técnicas descritas proporcionan secuenciación indexada, mediante la cual los fragmentos de ácido nucleico procedentes de una muestra particular se indexan utilizando un conjunto preestablecido de secuencias índice en cada sitio de índice. Aunque se describen ciertas realizaciones en el contexto de dos sitios de índice separados por cada fragmento de ácido nucleico (es decir, indexación doble), debe entenderse que las técnicas descritas pueden usarse con fragmentos de ácido nucleico que tienen al menos un sitio de índice, al menos dos sitios de índice y que pueden incorporar tres, cuatro o más sitios de índice para cada fragmento de ácido nucleico. Las muestras indexadas como se proporcionan en el presente documento se pueden combinar con otras muestras indexadas para secuenciación, siempre que las otras muestras indexadas estén indexadas con un conjunto diferente de secuencias índice en cada sitio de índice. En consecuencia, las presentes técnicas permiten reacciones de secuenciación agrupadas o de múltiples muestras mediante las cuales cada muestra individual en la reacción de secuenciación se indexa con un

conjunto único asignado de secuencias índice y las lecturas de secuenciación adquiridas se asocian con la muestra de origen adecuada a través de los índices únicos asignados.

Las presentes técnicas mejoran la precisión de la secuenciación con respecto a las técnicas que incorporan una única secuencia índice por fragmento de muestra y/o en cada sitio de índice. Por ejemplo, el análisis de secuenciación se puede realizar usando una asignación basada en reglas que descarta cualquier lectura de secuenciación que no incluya un índice asignado adecuado, por ejemplo, tanto en el primer como en el segundo sitio de índice para fragmentos con doble índice o en solo un sitio de índice para fragmentos con índice individual. Los índices asignados seleccionados para cada muestra están diseñados para ser diversos y distinguibles utilizando diversas técnicas de secuenciación. Por ejemplo, cada secuencia índice tiene variabilidad interna, variabilidad dentro de su conjunto asignado y variabilidad en relación con otras secuencias índice asignadas a otras muestras del conjunto. De esta manera, los errores del dispositivo de secuenciación en una lectura de base individual en el índice son menos probables porque la diversidad de los índices reduce las lecturas erróneas. En un ejemplo, si un ciclo de secuenciación tiene baja complejidad o pocas muestras, los datos de la secuencia índice pueden, no obstante, reflejar una diversidad mayor de la que se esperaría si el número de secuencias índice diferentes en cada sitio de índice tuviera una correspondencia 1:1 con el número de muestras. Es decir, en lugar de utilizar una sola secuencia índice por cada sitio de índice de muestra y donde cada muestra está asociada con solo un índice en cada sitio de índice, lo que produciría una correspondencia 1:1, cada muestra se asocia con múltiples secuencias índice potenciales en cada sitio de índice. Por consiguiente, las secuencias índice diferentes o distinguibles están presentes con respecto a un número de muestras en una relación de al menos 2:1, al menos 3:1 o al menos 4:1 para cada sitio de índice. En un ejemplo, para técnicas de secuenciación que incorporan lecturas de índices específicas (es decir, usando un cebador de índice dirigido justo cadena arriba de la secuencia índice y secuenciando un número limitado de bases correspondientes a la longitud de la secuencia índice), las técnicas presentes y la diversidad de índice mejorada puede dar lugar a lecturas de nucleótidos más precisas.

Al aumentar la precisión de los datos de las secuencias índice, se descartan menos lecturas de secuenciación según la asignación por reglas, lo que también mejora la precisión de la secuenciación incluso para muestras de baja concentración y en casos de bajo número de muestras. Además, la asociación de muestras mejorada mediante una mejora de la precisión de la secuenciación de índice reduce los efectos del salto de índices (es decir, la asignación errónea de una lectura de secuenciación a una muestra por la identificación errónea del índice).

Según un aspecto de la presente invención, se proporciona un método implementado por ordenador para asignar datos de secuenciación representativos de una secuencia de interés a una muestra, como se define en la reivindicación 1. Otras realizaciones se proporcionan como se define en las reivindicaciones dependientes.

Figuras

La FIG. 1 es una ilustración esquemática de un fragmento de ácido nucleico índice según las presentes técnicas;

la FIG. 2 es una ilustración esquemática de una biblioteca de fragmentos de ácido nucleico indexados según las presentes técnicas;

la FIG. 3 es un diagrama de flujo de métodos para generar una biblioteca de ácidos nucleicos indexados de una muestra según las presentes técnicas;

la FIG. 4 es un diagrama de flujo de métodos de secuenciación de bibliotecas de ácidos nucleicos indexados agrupados según las presentes técnicas;

la FIG. 5 es una ilustración esquemática de bibliotecas de fragmentos de ácido nucleico indexados agrupados según las presentes técnicas;

la FIG. 6 es una ilustración esquemática de adaptadores de secuenciación que incluyen secuencias índice según las presentes técnicas;

la FIG. 7 es un diagrama de flujo de métodos de secuenciación de una biblioteca de ácidos nucleicos según las presentes técnicas;

la FIG. 8 es un ejemplo de un índice establecido según las presentes técnicas;

la FIG. 9 es un ejemplo de un índice establecido según las presentes técnicas;

la FIG. 10 es un kit de secuenciación de ácidos nucleicos según las presentes técnicas;

la FIG. 11 es un diagrama de bloques de un dispositivo de secuenciación configurado para adquirir datos de secuenciación según las presentes técnicas; y

la FIG. 12 muestra un ejemplo de una interfaz gráfica de usuario según las presentes técnicas.

Descripción detallada

Las presentes técnicas están dirigidas a ácidos nucleicos indexados y a la secuenciación y el análisis mediante su uso. La secuenciación de ácidos nucleicos genera grandes cantidades de datos brutos que posteriormente se analizan y compilan para proporcionar información significativa relacionada con la muestra secuenciada. Las técnicas de secuenciación han avanzado hasta procesar múltiples muestras simultáneamente, lo que proporciona ahorro de tiempo y costes. Sin embargo, estas técnicas presentan desafíos en el procesamiento de datos. Los datos de secuenciación adquiridos en un ciclo de secuenciación de múltiples muestras deben asociarse con su muestra de origen antes de poder realizar el ensamblaje y el análisis de la secuencia para cada muestra individual. Sin embargo, a menudo dicha asociación de muestras no se puede completar con precisión utilizando únicamente las secuencias de ácidos nucleicos endógenas. En consecuencia, ciertas técnicas de secuenciación incorporan un código de barras exógeno único o una secuencia índice en los ácidos nucleicos antes de la secuenciación, estando asociada cada muestra con un código de barras o índice único. Una vez adquiridos los datos de secuenciación, las lecturas de secuencia que tienen el código de barras o índice único se asignan a la muestra de origen adecuada.

Si bien estas técnicas facilitan la asignación de datos de secuenciación de múltiples muestras, todavía se producen errores en la asignación de muestras a las lecturas de secuenciación. Independientemente de la fuente de los errores de asignación de muestras introducidos, las asignaciones erróneas de secuencias pueden dar lugar a imprecisiones en el ensamblaje posterior del genoma y/o en el análisis de datos que se realiza a partir de datos mal asignados. Si los datos de secuencias índice no se adquieren con precisión mediante un dispositivo de secuenciación, una lectura de secuenciación de un fragmento de ácido nucleico (p. ej., representativo de un fragmento de ácido nucleico de 50-300 pb) puede asignarse a la muestra incorrecta basándose en datos inexactos. Además, determinadas técnicas de secuenciación pueden estar asociadas con un mayor grado de salto de índices o recombinación molecular de índices entre muestras. El salto de índices puede deberse a que un extremo de esta molécula (incluida la región de índice) se transfiera entre moléculas en la biblioteca de moldes. Si bien el salto de índices puede producirse a una tasa baja (~0,05 %), incluso niveles bajos de salto de índices pueden ser relevantes en un contexto clínico. En el caso de una lectura de índice única, una transferencia dará lugar a la asignación incorrecta de una lectura de secuenciación a otra muestra, lo que a su vez provocará contaminación en el análisis posterior. El salto de índices también puede ocurrir con la indexación doble mediante enfoque combinatorio (p. ej., asignando la misma secuencia a múltiples muestras en el índice uno o en el índice dos, pero garantizando que cualquier combinación particular de índice uno e índice dos sea exclusiva de una muestra específica). En determinadas realizaciones, las presentes técnicas se utilizan junto con índices exclusivos dentro de la primera y segunda lecturas de índice. En ese caso, incluso si se produce un intercambio, la combinación de índices observada no pertenecerá al conjunto esperado de pares de índices.

En el presente documento se proporcionan técnicas de indexación que dan lugar a una reducción de la asignación errónea de lecturas de secuenciación. A diferencia de las técnicas en las que cada muestra está asociada con un único código de barras o índice, las presentes técnicas proporcionan una pluralidad de secuencias índice que están asociadas de forma exclusiva con cada muestra individual y que se introducen durante la preparación de la muestra. La pluralidad de secuencias índice se introduce en el sitio de índice (o sitios de índice), de modo que los fragmentos de ácido nucleico individuales preparados a partir de una muestra tengan al menos una secuencia índice y de manera que todas las diferentes secuencias índice estén presentes en una biblioteca de los fragmentos de ácido nucleico de cada muestra. De esta manera, se introduce diversidad de secuencias índice muestra a muestra. Esta diversidad puede mejorar la precisión de los datos de secuencia de índice adquiridos. En particular, los dispositivos de secuenciación que adquieren datos de imágenes representativos de cientos (o miles) de fragmentos de ácido nucleico de dos o más muestras que se secuencian simultáneamente pueden tener dificultades para determinar las identificaciones de nucleótidos cuando las muestras carecen de diversidad de nucleótidos, lo que a su vez puede dar lugar a datos de imágenes adquiridos difíciles de resolver. Por ejemplo, determinadas técnicas de secuenciación evalúan las diferencias en la intensidad de la señal de la imagen para realizar identificaciones de nucleótidos. Para las muestras que solo tienen una única secuencia índice por cada sitio de índice, los nucleótidos en cada ciclo de secuenciación de la secuencia índice son tan diversos como el número total de muestras y pueden no ser lo suficientemente diversos para ciclos con bajos números de muestras. En consecuencia, en una realización, las técnicas de indexación proporcionadas en el presente documento dan lugar a una adquisición mejorada de datos de secuencias índice a través de la diversidad de índices y permiten una secuenciación precisa incluso para ciclos de secuenciación de baja multiplicidad.

Con ese fin, la FIG. 1 es una ilustración esquemática de un fragmento 12 de ácido nucleico indexado de una muestra según las presentes técnicas. El fragmento 12 de ácido nucleico indexado representa un fragmento adecuado para un ciclo de secuenciación. El fragmento 12 de ácido nucleico indexado incluye una inserción 14 de ácidos nucleicos procedentes de la muestra, es decir, ácidos nucleicos endógenos. El fragmento 12 de ácido nucleico indexado también incluye secuencias introducidas o exógenas que facilitan la secuenciación. Dichas secuencias pueden incluir una o más secuencias de cebadores de secuenciación 16, 18 que están en 5' y 3' de la inserción 14 y que permiten la unión de cebadores de secuenciación universales a una o ambas cadenas. El fragmento 12 de ácido nucleico indexado también incluye un primer sitio 20 de índice y un segundo sitio 22 de índice. En la realización ilustrada, el fragmento 12 de ácido nucleico indexado también incluye una primera secuencia 26 adaptadora y una segunda secuencia 28

adaptadora. La secuencia o secuencias 26, 28 adaptadoras se pueden seleccionar basándose en la plataforma de secuenciación deseada y pueden ser, por ejemplo, adaptadores P7 y P5 como se ilustra, que facilitan la unión a la celda de flujo o al sustrato de secuenciación del fragmento 12 de ácido nucleico indexado.

El fragmento 12 de ácido nucleico indexado ilustrado puede ser un fragmento bicatenario y la primera secuencia 20 índice puede ubicarse en 5' de la inserción 14 y la segunda secuencia 22 índice puede ubicarse en 3' de la inserción 14 en la cadena directa de manera que las secuencias 20, 22 índice flanqueen la inserción. La primera secuencia 20 índice se selecciona de un primer conjunto 30 de índices, que incluye una pluralidad (p. ej., dos, tres, cuatro o más) secuencias índice, ilustradas como secuencias 20a, 20b, 20c, 20d índice. La segunda secuencia 22 índice, cuando está presente, se selecciona de una segunda secuencia índice, que incluye una pluralidad de secuencias índice, ilustradas como secuencias 22a, 22b, 22c, 22d índice.

Dentro del conjunto de índices (p. ej., el conjunto 30 de índices o el conjunto 32 de índices), las secuencias índices individuales (p. ej., la secuencia 20a, 20b, 20c, 20d índice) son diferentes entre sí. Es decir, tienen secuencias diferentes entre sí, como se establece en el presente documento. Además, las secuencias índice en el primer conjunto 30 de índices también se distinguen de las secuencias índice en el segundo conjunto 32 de índices. En determinadas realizaciones, las secuencias 20, 22 índice se pueden distinguir de los ácidos nucleicos endógenos de la muestra o muestras. Sin embargo, en otras realizaciones, las secuencias índice pueden no distinguirse necesariamente de la inserción 14.

Las secuencias índice pueden ser monocatenarias o bicatenarias y pueden tener al menos 5 bases, al menos 6 bases, al menos 7 bases, al menos 8 bases o más. En determinadas realizaciones, las secuencias índice tienen una longitud de 5-8 bases, 5-10 bases, 5-15 bases, 5-25 bases, 8-10 bases, 8-12 bases, 8-15 bases u 8-25 bases, etc. Además, en determinadas realizaciones, las secuencias índice (p. ej., secuencias 20, 22 índice) tienen una longitud de no más de 30 bases, no más de 25 bases, no más de 20 bases, no más de 15 bases. Debe entenderse que la longitud de las secuencias índice como se proporciona en el presente documento puede hacer referencia a las partes únicas/distinguibles de las secuencias dentro de cada conjunto de índices (y en relación con otros conjuntos de índices) y puede excluir secuencias comunes o universales adyacentes del fragmento 12 de ácido nucleico indexado que puede servir como cebadores de secuenciación y que son comunes a todos los fragmentos 12 de ácido nucleico indexados de una muestra.

Si bien ciertas realizaciones de las técnicas descritas se analizan en el contexto de las técnicas de secuenciación de índice doble, debe entenderse que las técnicas proporcionadas en el presente documento también se pueden usar en el contexto de la secuenciación de índice individual. Por ejemplo, un fragmento 12 de ácido nucleico puede incorporar solo una secuencia índice (p. ej., la secuencia 20 o 22 índice) seleccionada de un conjunto de índices (p. ej., el conjunto 30 de índices). Además, las secuencias índice pueden seleccionarse para secuenciarse a partir de una o ambas cadenas de un fragmento 12 de ácido nucleico bicatenario, p. ej., en secuenciación de lectura única o de ambos extremos, dependiendo de la técnica de secuenciación deseada.

La FIG. 2 es una ilustración esquemática de una biblioteca 40 de secuenciación procedente de una muestra y que incluye fragmentos 12 de ácido nucleico indexados. Como se ilustra, la biblioteca 40 tiene varias configuraciones diferentes de combinaciones de índices. Por ejemplo, ciertos fragmentos 12a pueden indexarse con la combinación del índice 20a y el índice 22a, mientras que otros fragmentos 12b pueden indexarse con la combinación de los índices 20b y 22a. Se contempla que, durante la preparación de la biblioteca 40 a partir de la muestra, se utilicen mezclas de adaptadores que incluyan las diferentes secuencias 20, 22 índice en cada conjunto de índices (p. ej., el primer conjunto 30 de índices y el segundo conjunto 32 de índices) para modificar las inserciones 14 de una manera generalmente aleatoria de modo que cualquier inserción 14 dada pueda modificarse con una cualquiera de una secuencia 20a, 20b, 20c o 20d índice individual del primer conjunto 30 de índices y una cualquiera de la secuencia 22a, 22b, 22c o 22d índice individual del segundo conjunto 32 de índices. La FIG. 2 muestra una parte de las posibles combinaciones de la primera secuencia 20 índice y la segunda secuencia 22 índice que están presentes en la biblioteca. Además, debe entenderse que son posibles combinaciones adicionales, dependiendo del número total de secuencias índice en un conjunto de índices particular usado para la indexación. También debe entenderse que, para una muestra particular, múltiples fragmentos 12 de ácido nucleico indexados con respectivas inserciones 14 diferentes tendrán la misma configuración o combinación de secuencias 20, 22 índice.

La FIG. 3 es un diagrama de flujo de una realización de un método 50 para preparar la biblioteca 40 indexada de la FIG. 2. Después de adquirir la muestra de interés (etapa 52), los ácidos nucleicos de la muestra se fragmentan (etapa 54). Los ácidos nucleicos fragmentados están en contacto con una pluralidad de secuencias índice diversas de un conjunto de índices. En particular, al inicio de la preparación de la biblioteca, se pueden asociar o asignar uno o más conjuntos de índices a una muestra particular. La preparación de la biblioteca puede entonces prepararse utilizando solo un conjunto o conjuntos de índices asignados a muestras (etapa 56). En realizaciones particulares, la pluralidad de secuencias índice asignadas a muestras, p. ej., el primer conjunto 30 de índices o el segundo conjunto 32 de índices, se proporciona con concentraciones iguales de cada secuencia índice individual, p. ej., la secuencia 20, 22 índice, en el conjunto. Es decir, si el conjunto 30 de índices tiene tres secuencias 20a, 20b, 20c índice diferentes, se proporcionan en una relación 1:1:1 entre sí y, en una realización, se proporcionan como una mezcla que tiene

concentraciones iguales (o concentraciones aproximadamente iguales en un intervalo de 0,9 a 1,1 entre sí) de cada secuencia 20 índice.

Los ácidos nucleicos fragmentados se modifican con los adaptadores que incluyen las secuencias índice asignadas a muestras (etapa 58) (p. ej., como se muestra en la FIG. 4). En otras realizaciones, las secuencias índice se pueden añadir a los ácidos nucleicos fragmentados en una etapa separada de la adición de los adaptadores. Como resultado de la modificación, se generan fragmentos de ácido nucleico indexados. Proporcionar las secuencias índice individuales de cada conjunto de índices en concentraciones sustancialmente iguales puede facilitar la incorporación relativamente igual de cada secuencia índice individual dentro de los fragmentos 12 de ácido nucleico indexados.

La FIG. 4 es una ilustración esquemática de una realización de la preparación de una biblioteca de fragmentos 12 de ácido nucleico índice de una muestra 60. Sin embargo, debe entenderse que el método ilustrado es a modo de ejemplo y que los ácidos nucleicos 12 indexados como se proporcionan en el presente documento pueden prepararse usando otras técnicas de preparación de bibliotecas, p. ej., tagmentación. En determinadas realizaciones, la biblioteca se utiliza para obtener moléculas molde monocatenarias que se pueden utilizar en una reacción de secuenciación. La biblioteca puede formarse a partir de fragmentos de ácidos nucleicos con secuencias comunes en sus extremos 5' y 3' pero con diversidad en uno o más sitios de índice y en un sitio de inserción. Como se explica con más detalle a continuación, los fragmentos de ácido nucleico dentro de la biblioteca pueden contener regiones de secuencia común en (o próximas a) sus extremos 5' y 3'. En determinadas realizaciones, los fragmentos de ácido nucleico de la biblioteca están "bifurcados", de modo que una secuencia común en el extremo 5' de cada molde individual en la biblioteca no es idéntica y no es completamente complementaria de la secuencia común en el extremo 3' de dicho molde. Sin embargo, en otras realizaciones, los adaptadores no están bifurcados, p. ej., son completamente complementarios.

La muestra 60 se fragmenta para generar ácidos nucleicos 62 fragmentados con extremos salientes que se hacen romos mediante varios métodos conocidos por los expertos en la técnica. En un método, los extremos del ADN fragmentado se reparan en los extremos con ADN polimerasa T4 y polimerasa Klenow y a continuación se fosforilan con una enzima polinucleótido cinasa para generar ácidos nucleicos 64 fragmentados fosforilados. A continuación, se añade un solo desoxinucleótido "A" en ambos extremos 3' de las moléculas de ADN utilizando la enzima Taq polimerasa, que produce ácidos nucleicos 68 fragmentados con un saliente de una base en 3' que son complementarios del otro saliente "T" de una base en 3' en el extremo bicatenario del adaptador bifurcado.

A continuación, se realiza una reacción de ligadura entre el adaptador 70 bifurcado y los ácidos nucleicos 68 fragmentados con saliente de una base en 3' utilizando una enzima ligasa adecuada (p. ej., ADN ligasa T4) que une dos copias del adaptador a cada fragmento de ADN, uno en cada extremo, para formar construcciones 72 de adaptador-diana. Los productos de esta reacción pueden purificarse a partir de adaptador no ligado mediante varias técnicas, incluyendo cromatografía de inclusión por tamaño, preferiblemente por electroforesis a través de un gel de agarosa seguido de escisión de una parte de la agarosa que contiene el ADN de mayor tamaño que el tamaño del adaptador.

En particular, en la realización representada, los adaptadores 70 bifurcados implementados en la preparación de la biblioteca incluyen diversas mezclas de índices en la primera secuencia 20 índice y la segunda secuencia 22 índice. A diferencia de otras técnicas en las que los adaptadores acoplados a la muestra o a la inserción de ácidos nucleicos son generalmente idénticos, los adaptadores 70 bifurcados se preparan usando una mezcla de las secuencias índice del conjunto 30 de índices y del conjunto 32 de índices, por ejemplo. En consecuencia, los adaptadores 70 bifurcados no son todos idénticos entre sí y son diversos en función de la combinación particular de secuencias índice en la primera secuencia 20 índice y la segunda secuencia 22 índice. Sin embargo, los cebadores 16, 18 de secuenciación y las secuencias adaptadoras son compartidos o comunes entre los adaptadores 70 bifurcados. Las construcciones 72 de adaptador-diana pueden amplificarse para generar los fragmentos 12 de ácido nucleico indexados, que, a su vez, pueden desnaturalizarse para separar la estructura bicatenaria en moléculas monocatenarias antes de la secuenciación.

Puede prepararse una biblioteca 40 de fragmentos 12 de ácido nucleico indexados a partir de una única muestra y agruparse con otras bibliotecas de otras muestras respectivas y prepararse de manera similar, según la plataforma de secuenciación deseada. La FIG. 5 es una ilustración esquemática de bibliotecas que tienen diferentes conjuntos 30, 32 de índices específicos o asignados que pueden agruparse y distinguirse después de la secuenciación a través de las secuencias índice distinguibles presentes en los primeros conjuntos 30a, 30b, 30c de índices y/o las secuencias índice distinguibles presentes en los segundos conjuntos 32a, 32b, 32c de índices. El primer conjunto 30a de índices, utilizado en la creación de la biblioteca 40a, incluye un conjunto distinto de secuencias de índices que no están presentes y no solapan con otros conjuntos (30b, 32b, 30c, 32c) de índices presentes en otras bibliotecas 40b, 40c. También debe entenderse que el primer conjunto 30a de índices también es distinto y no solapa con el segundo conjunto 32a de índices utilizado en la misma biblioteca 40a.

La FIG. 6 es un diagrama de flujo de un método 80 de secuenciación de bibliotecas agrupadas de diferentes muestras (p. ej., las bibliotecas de la FIG. 5) y el uso de las secuencias índice proporcionadas en el presente documento para asignar datos de secuenciación a la muestra correcta. En la etapa 82, se preparan fragmentos de muestras individuales indexados (p. ej., fragmentos 12 de ácido nucleico indexados) a partir de una muestra individual usando al menos un

conjunto de índices que comprende una pluralidad de secuencias índice. Los fragmentos de muestras individuales indexados se agrupan con otros fragmentos de muestras indexados de diferentes muestras preparadas usando conjuntos de índices diferentes (es decir, distinguibles de los índices a los que se hace referencia en la etapa 82) para generar fragmentos de muestras combinados en la etapa 84. El agrupamiento o la combinación puede tener lugar durante la carga de la muestra sobre un sustrato de secuenciación. En una realización, las muestras combinadas se cargan en diferentes carriles de una celda de flujo. Debido a que los carriles de la celda de flujo están separados entre sí durante una secuenciación, los conjuntos de índices utilizados en las muestras en un primer carril de la celda de flujo se pueden usar para indexar otras muestras en carriles diferentes, siempre que no se indexen muestras dentro del mismo carril utilizando los mismos conjuntos de índices.

En la etapa 86, se adquieren datos de secuenciación que son representativos de la secuenciación de los fragmentos de muestras combinados y, en la etapa 88, las lecturas de secuenciación se asocian con la muestra individual solo cuando las lecturas de secuenciación incluyen una secuencia índice del conjunto de índices asignado. En determinadas realizaciones, cuando los fragmentos de muestras indexados tienen cada uno una primera secuencia índice y una segunda secuencia índice, una asignación por reglas requiere que tanto la primera secuencia índice como una segunda secuencia índice sean miembros de los conjuntos de índices asignados para una muestra determinada. El fallo en un solo sitio de índice es suficiente para dar lugar a la eliminación de la lectura de secuenciación del análisis posterior (p. ej., ensamblaje del genoma) para eliminar lecturas con salto de índices. En determinadas realizaciones, la lectura de secuenciación fallida puede almacenarse para evaluación de la calidad. Es decir, determinadas secuencias índice potenciales pueden estar asociadas con un mayor fallo del índice y se puede realizar un seguimiento de ellas para su rediseño.

La FIG. 7 es una ilustración esquemática de una realización de una técnica 90 de secuenciación utilizada para adquirir datos de secuenciación como se proporciona en el presente documento. Como se ilustra, los ácidos nucleicos secuenciados se inmovilizan sobre un sustrato 92 mediante una sonda 94 de captura complementaria de la utilizada junto con una cadena 96 molde procedente de la desnaturalización de los fragmentos 12 de ácido nucleico indexados. La primera lectura de secuenciación, lectura 1, es una secuencia de la inserción 14 que se adquiere mediante contacto con un cebador 98 de lectura 1 que se dirige a (es decir, es complementario de) una de las secuencias 16, 18 del cebador de secuenciación. Por ejemplo, si la cadena 96 molde se captura usando la sonda 94 de captura complementaria de un adaptador p7 (u otro 5'), el cebador 98 de lectura 1 puede ser complementario de la secuencia 18 del cebador de secuenciación.

Después de retirar el producto de lectura 1, la técnica 90 también puede adquirir una primera lectura de índice de la secuencia 20 índice en el lado p7 de la inserción usando un primer cebador 100 de índice, que puede dirigirse al cebador 16 de secuenciación. En la realización representada, la primera lectura del índice está en la misma cadena que la lectura 1. Si bien la lectura 1 puede ser de 100 a 150 bases, las lecturas de índices pueden ser relativamente más cortas, p. ej., de 8 a 12 bases, o tan largas como una longitud conocida de las secuencias 20, 22 índice. De esta manera, se conservan los recursos de secuenciación. Para adquirir una segunda lectura de índice de la misma cadena después de la retirada del primer producto de índice, se puede usar la cadena molde capturada en el extremo 3' mediante una sonda 104 de captura. Por ejemplo, la segunda lectura de índice puede adquirirse con un segundo cebador 106 de lectura de índice dirigido a una parte del adaptador p5 (u otro 3') o secuencia adyacente. Sin embargo, en otras realizaciones, la segunda lectura de índice puede adquirirse a partir de una cadena complementaria resintetizada. Después de la retirada del producto de la segunda lectura de índice, se sintetiza una cadena 110 complementaria de la cadena 96 molde y se retira la cadena 96 molde original. Posteriormente, la cadena sintetizada se pone en contacto con un cebador 112 de lectura 2 para obtener una secuencia de lectura 2 que es un complemento inverso de la secuencia 1 de lectura. Se contempla que el primer cebador 100 de índice, el segundo cebador 106 de índice y los cebadores 98, 112 de lectura 1 y lectura 2 son universales para todas las cadenas 96 molde, independientemente del origen de la muestra de molde y la secuencia índice. En función de las lecturas del primer índice y del segundo índice adquiridas, los datos de secuenciación de lectura 1 y lectura 2 pueden asociarse con una muestra particular.

Como se proporciona en el presente documento, los conjuntos de índices proporcionan diversidad adicional para una muestra determinada en la lectura del índice. Al asignar una única secuencia índice a cada muestra, existe el riesgo de que no todos los nucleótidos estén representados en un ciclo determinado de la lectura del índice cuando el número de muestras es pequeño. Cuando se hace esto, es posible que el ciclo de secuenciación no genere datos utilizables porque los algoritmos de corrección de intensidad e identificación de nucleótidos no funcionan según lo diseñado en los ciclos de índice, lo que da lugar a una falta de asignación de las lecturas a las muestras. El ejemplo más extremo de esta circunstancia es la secuenciación de una sola muestra, lo que, por ejemplo, se puede realizar para un seguimiento positivo de la muestra y/o para eliminar Phi X no indexado de las lecturas. Una solución a este problema es marcar cada muestra con múltiples secuencias índice de modo que haya suficiente representación de nucleótidos para que el proceso de análisis primario funcione según lo diseñado. La FIG. 8. es una representación de un conjunto de índices (p. ej., conjunto 30 de índices) con nucleótidos N₁, N₂, etc. El conjunto de índices puede seleccionarse de manera que, para cualquier posición dada, los nucleótidos sean diversos dentro de las secuencias del conjunto de índices. Por ejemplo, en la posición 120, los nucleótidos pueden ser diferentes de modo que tres o más de A, C, T y G estén representados en todo el conjunto de índices. En la FIG. 9, que representa una realización específica de un conjunto 30 de índices, se representan tres nucleótidos diferentes en la posición 120 y se representan cuatro

nucleótidos diferentes en la posición 122. En consecuencia, los conjuntos de índices pueden diseñarse de manera que, en cada posición de nucleótido, estén representados al menos tres nucleótidos diferentes en todo el conjunto de índices. En otras realizaciones, se representan cuatro nucleótidos diferentes a lo largo del conjunto de índices en al menos la mitad de las posiciones de nucleótidos en la secuencia índice. Además, en determinadas realizaciones, las secuencias índice son internamente diversas. Es decir, al menos una parte de los nucleótidos adyacentes de una secuencia índice individual son diferentes entre sí.

Además, es posible que determinados índices no funcionen tan bien como otros, lo que da lugar a que determinadas muestras estén subrepresentadas incluso si la concentración de las muestras introducidas es idéntica. Una ventaja de colocar múltiples secuencias índice en cada muestra es que el impacto general del rendimiento deficiente de una secuencia índice es limitado. Dentro de esta estrategia, también es posible agrupar secuencias índice de alto y bajo rendimiento para mejorar aún más la uniformidad de la representación entre diferentes muestras.

La tabla 1 y la tabla 2 son ejemplos de conjuntos de índices (p. ej., el conjunto 30, 32 de índices) y las secuencias de índices individuales (p. ej., la secuencia 20, 22 índice) que forman un conjunto de índices según las presentes técnicas. Por ejemplo, las secuencias índice (p. ej., secuencias 20, 22 índice) bajo cada número de grupo individual se contemplan como un conjunto de índices único. Es decir, el grupo n.º 0 incluye cuatro secuencias F7-001, F7-002, F7-003 y F7-004 índice diferentes que juntas forman un único conjunto de índices. El grupo n.º 1 incluye cuatro secuencias F7-005, F7-006, F7-007 y F7-008 índice diferentes adicionales que juntas forman un único conjunto de índices, y así sucesivamente.

En los conjuntos de índices que se muestran a continuación, hay cuatro índices por cada conjunto de índices. Sin embargo, debe entenderse que el tamaño del conjunto de índices puede variar y puede incluir tres, cuatro o más secuencias índice distinguibles individuales. Los conjuntos de índices mostrados se seleccionaron para garantizar que dentro de cualquier grupo haya una representación equilibrada de nucleótidos. Específicamente, dentro de cualquier grupo, el segundo nucleótido menos frecuente en cada ciclo debe estar presente en al menos el 25 % de las secuencias índice.

La tabla 1 muestra conjuntos de índices que pueden ser un conjunto 30 de índices del lado P7. Sin embargo, en determinadas realizaciones, las secuencias de la tabla 1 se pueden usar en el lado P5 como el conjunto 32 de índices.

Tabla 1

Conjuntos de índices P7	
Grupo de índices n.º 0	
F7-001,	CCATACTA
F7-002,	TGTGCGCT
F7-003,	CACATTGC
F7-004,	ATCCGGAG
Grupo de índices n.º 1	
F7-005,	ACCTTAAC
F7-006,	CAGCGCCT
F7-007,	TGAACAGG
F7-008,	GTGGCTCA
Grupo de índices n.º 2	
F7-009,	GGTAACAC
F7-010,	TTCTGATG
F7-011,	ACGCTTGT
F7-012,	GAATAGCA
Grupo de índices n.º 3	
F7-013,	GGATTCAA
F7-014,	ATTGAACT
F7-015,	TCCAAGGC
F7-016,	TAACCTTG

ES 2 966 831 T3

	Conjuntos de índices P7
	Grupo de índices n.º 0
5	Grupo de índices n.º 4
	F7-017, GAGACAAC
	F7-018, TGTTAGGA
	F7-019, ACAGTATG
10	F7-020, CTTCTTCT
	Grupo de índices n.º 5
	F7-021, GTCGCCTT
15	F7-022, CGGTGAGA
	F7-023, TCACTGAA
	F7-024, AGCAATTG
20	Grupo de índices n.º 6
	F7-025, TAGGTTGA
	F7-026, CCTTCGCC
	F7-027, GGATAATG
25	F7-028, TTCACAAT
	Grupo de índices n.º 7
	F7-029, TGGCACGG
30	F7-030, ATATCTAC
	F7-031, GCTCGGTT
	F7-032, CTAATGTA
35	Grupo de índices n.º 8
	F7-033, AAGATGAA
	F7-034, GCCGAATC
40	F7-035, AGATGCGG
	F7-036, TTCCATCC
	Grupo de índices n.º 9
45	F7-037, CCTCACGT
	F7-038, GTGTGGAC
	F7-039, TCCGTTCC
	F7-040, TGAGGATA
50	Grupo de índices n.º 10
	F7-041, TGATTGCC
	F7-042, AACAGCTT
55	F7-043, CATCATAA
	F7-044, TTGTCAGC
	Grupo de índices n.º 11
60	F7-045, CGGAACTT
	F7-046, GTACTTGG
	F7-047, AACGCACC
	F7-048, GCTTAGAG
65	Grupo de índices n.º 12

	Conjuntos de índices P7
	Grupo de índices n.º 0
5	F7-049, TAATCGAT
	F7-050, CCGGTCCA
	F7-051, ATTAGTTC
	F7-052, CGCGAAGG
10	Grupo de índices n.º 13
	F7-053, CTCTAGCT
	F7-054, TCGATCTG
15	F7-055, AATGGAGA
	F7-056, CGGCCTAT
	Grupo de índices n.º 14
20	F7-057, TGCCTCTT
	F7-058, CCAGCTGC
	F7-059, AAGTGTCA
25	F7-060, GTTATACG
	Grupo de índices n.º 15
	F7-061, GTAGAGGC
	F7-062, TAGTCCTA
30	F7-063, CCTAGAAT
	F7-064, GACCTCCG

La tabla 2 muestra conjuntos de índices que pueden ser un conjunto 32 de índices del lado P5. En determinadas realizaciones, los conjuntos de índices están diseñados para usarse junto con el conjunto 30 de índices P7 que tiene el mismo número de grupo para generar fragmentos de ácido nucleico indexados para una muestra. Por ejemplo, determinados conjuntos emparejados pueden haber completado la evaluación de calidad cuando se usan juntos y pueden estar asociados con niveles bajos de salto de índices o rendimientos de amplificación generalmente distribuidos de forma uniforme. En otras realizaciones, cualquier conjunto 30 de índices del lado P7 (o del lado 5') dado puede usarse con cualquier otro conjunto 32 de índices del lado P5.

	Conjuntos de índices P5
	Grupo de índices n.º 0
45	F5-001, TCGCTCTA
	F5-002, ATTGGAGG
	F5-003, AACTAGAC
50	F5-004, CGGACTAT
	Grupo de índices n.º 1
	F5-005, TCCTTAGG
55	F5-006, AGGAGGAA
	F5-007, CAACACTC
	F5-008, GTGGTTCT
	Grupo de índices n.º 2
60	F5-009, TGGTGGTT
	F5-010, GCTACCGC
	F5-011, AGCGTTCA
65	F5-012, CAATTATG

ES 2 966 831 T3

	Conjuntos de índices P5
	Grupo de índices n.º 0
5	Grupo de índices n.º 3
	F5-013, TTACAAGA
	F5-014, GGTTCTAC
10	F5-015, ACGTGGCG
	F5-016, CACAATTG
	Grupo de índices n.º 4
15	F5-017, CCATTCAA
	F5-018, TAGCAGGC
	F5-019, TTCAGATG
	F5-020, AGTGCGGT
20	Grupo de índices n.º 5
	F5-021, ATTCAACT
	F5-022, TCAAGGAG
25	F5-023, CACCTCGA
	F5-024, AGGTATCC
	Grupo de índices n.º 6
30	F5-025, AAGCGCTT
	F5-026, TGTGTAGC
	F5-027, CTATAGAG
	F5-028, GCCACTTA
35	Grupo de índices n.º 7
	F5-029, TGCCTTG
	F5-030, AACATGGT
40	F5-031, GTGTGCAA
	F5-032, CCTGAACA
	Grupo de índices n.º 8
45	F5-033, CTTCTTGC
	F5-034, TAGGAATA
	F5-035, AGTTGCCT
	F5-036, TCCACAAT
50	Grupo de índices n.º 9
	F5-037, AGACCTCT
	F5-038, CTCATGTA
55	F5-039, CCGTACGC
	F5-040, GAAGGTCG
	Grupo de índices n.º 10
60	F5-041, AGATTGAT
	F5-042, TCTGATTC
	F5-043, ATCCGCCA
	F5-044, GAGACGAG
65	Grupo de índices n.º 11

	Conjuntos de índices P5
	Grupo de índices n.º 0
5	F5-045, CGCTAACT
	F5-046, ATAGCTAG
	F5-047, GCACTGGC
10	F5-048, TCGAGCGT
	Grupo de índices n.º 12
	F5-049, GGCAAGTT
	F5-050, CTGTTCCG
15	F5-051, ACCGGTGC
	F5-052, GATCCATA
	Grupo de índices n.º 13
20	F5-053, TTCACTCC
	F5-054, ACTTACTA
	F5-055, GAAGCAAT
25	F5-056, GGTCGTGA
	Grupo de índices n.º 14
	F5-057, GCCGTCAT
30	F5-058, TGTCGGCC
	F5-059, ATGAATGG
	F5-060, CATGGATT
	Grupo de índices n.º 15
35	F5-061, AGCCTAAG
	F5-062, CAAGCTGA
	F5-063, GTGTAATT
40	F5-064, TGTAACAC

La FIG. 10 es un ejemplo de un kit 150 de preparación de muestras para secuenciación que puede usarse junto con las presentes técnicas, es decir, para preparar fragmentos 12 de ácido nucleico indexados a partir de una muestra individual y, en determinadas realizaciones, para secuenciar los fragmentos 12 de ácido nucleico indexados. El kit 150 de preparación de muestras puede incluir un primer conjunto 30 de índices y, cuando se usa, un segundo conjunto 32 de índices. En ciertas realizaciones, el primer conjunto 30 de índices y/o el segundo conjunto 32 de índices se pueden proporcionar en forma de ácidos nucleicos adaptadores que incluyen elementos adicionales, tales como secuencias de cebadores, secuencias adaptadoras, etc. El primer conjunto 30 de índices y/o el segundo conjunto 32 de índices pueden proporcionarse dentro de respectivos recipientes individuales en cantidades premezcladas de modo que cada secuencia índice individual esté presente en concentraciones aproximadamente iguales y de modo que la diversidad de índices para una muestra determinada no sea tan susceptible a errores del usuario. En realizaciones en las que los adaptadores son adaptadores bifurcados de doble índice, ambos conjuntos 30, 32 de índices pueden estar presentes en un único ácido nucleico adaptador, que puede proporcionarse en un único recipiente. El kit 150 de preparación de muestras también puede incluir los cebadores 152 adecuados para usar junto con la plataforma de secuenciación deseada. El kit de preparación 150 de muestras también puede incluir una o más enzimas, tampones y/o reactivos 154 de preparación de muestras. El kit 150 de preparación de muestras puede proporcionarse como un kit preenvasado para preparar una biblioteca a partir de una única muestra o, en determinadas realizaciones, puede proporcionarse como un kit de múltiples muestras con una pluralidad de conjuntos 30, 32 de índices diferentes.

La FIG. 11 es un diagrama esquemático de un dispositivo 160 de secuenciación que puede usarse junto con las realizaciones descritas para adquirir datos de secuenciación de ácidos nucleicos indexados (p. ej., lecturas de secuenciación, lectura 1, lectura 2, lecturas de índice, lectura 1 de índice, lectura 2 de índice, datos de secuenciación de múltiples muestras) que se asignaron a muestras individuales utilizando las técnicas de indexación proporcionadas en el presente documento. El dispositivo 160 de secuenciación puede implementarse según cualquier técnica de secuenciación, tales como las que incorporan métodos de secuenciación por síntesis descritos en las publicaciones de patentes estadounidenses n.º 2007/0166705; 2006/0188901; 2006/0240439; 2006/0281109; 2005/0100900;

patentes estadounidenses n.º 7.057.026; WO 05/065814; WO 06/064199; WO 07/010.251. Alternativamente, se puede usar secuenciación mediante técnicas de ligadura en el dispositivo 160 de secuenciación. Tales técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de tales oligonucleótidos y se describen en la patente estadounidense n.º 6.969.488; patentes estadounidenses n.º 6.172.218; y la patente estadounidense n.º 6.306.597. Algunas realizaciones pueden utilizar secuenciación de nanoporos, mediante la cual las cadenas de ácido nucleico diana, o los nucleótidos eliminados de forma exonucleolítica de los ácidos nucleicos diana, pasan a través de un nanoporo. Cuando los ácidos nucleicos diana pasan a través del nanoporo, cada tipo de base puede identificarse midiendo las fluctuaciones en la conductancia eléctrica del poro (patente estadounidense n.º 7.001.792; Soni y Meller, Clin. Chem. 53, 1996-2001 (2007); Healy, Nanomed. 2, 459-481 (2007); y Cockcroft y col. J. Am. Chem. Soc. 130, 818-820 [2008]). Otras realizaciones más incluyen la detección de un protón liberado tras la incorporación de un nucleótido en un producto de extensión. Por ejemplo, la secuenciación basada en la detección de protones liberados puede usar un detector eléctrico y técnicas asociadas que son comercializados por Ion Torrent (Guilford, CT, una subsidiaria de Life Technologies) o métodos y sistemas de secuenciación descritos en los documentos US 2009/0026082 A1; US 2009/0127589 A1; US 2010/0137143 A1; o US 2010/0282617 A1. Algunas realizaciones particulares pueden utilizar métodos que suponen la monitorización en tiempo real de la actividad de la ADN polimerasa. Las incorporaciones de nucleótidos pueden detectarse mediante interacciones de transferencia de energía por resonancia de fluorescencia (FRET) entre una polimerasa portadora de fluoróforo y nucleótidos marcados con γ -fosfato o con guías de ondas de modo cero como se describe, por ejemplo, en Levene y col. Science 299, 682-686 (2003); Lundquist y col. Opt. Lett. 33, 1026-1028 (2008); Korch y col. Proc. Natl. Acad. Sci. EE. UU. 105, 1176-1181 (2008). Otras técnicas alternativas adecuadas incluyen, por ejemplo, secuenciación fluorescente localizada (FISSEQ) y secuenciación masiva de firmas en paralelo (MPSS). En realizaciones particulares, el dispositivo 160 de secuenciación puede ser un HiSeq, MiSeq o HiScanSQ de Illumina (La Jolla, CA). En otra realización, el dispositivo 160 de secuenciación puede configurarse para funcionar usando un sensor CMOS con nanopocillos fabricados sobre fotodiodos de modo que el depósito de ADN esté alineado uno a uno con cada fotodiodo.

El dispositivo 160 de secuenciación puede ser un dispositivo de detección de "un canal", en el que solo dos de cuatro nucleótidos están marcados y son detectables para cualquier imagen dada. Por ejemplo, la timina puede tener un marcador fluorescente permanente, mientras que la adenina usa el mismo marcador fluorescente en una forma desmontable. La guanina puede estar permanentemente oscura y la citosina puede estar inicialmente oscura pero se le puede añadir un marcador durante el ciclo. En consecuencia, cada ciclo puede implicar una imagen inicial y una segunda imagen en la que el colorante se escinde de cualquier adenina y se añade a cualquier citosina de modo que solo la timina y la adenina sean detectables en la imagen inicial pero solo la timina y la citosina sean detectables en la segunda imagen. Cualquier base que sea oscura en ambas imágenes es guanina y cualquier base que sea detectable en ambas imágenes es timina. Una base que sea detectable en la primera imagen pero no en la segunda es adenina y una base que no sea detectable en la primera imagen pero sí en la segunda es citosina. Al combinar la información de la imagen inicial y la segunda imagen, las cuatro bases pueden diferenciarse usando un canal.

En la realización representada, el dispositivo 160 de secuenciación incluye un dispositivo 162 de procesamiento de muestras separado y un ordenador 164 asociado. Sin embargo, como se ha observado, estos pueden implementarse como un solo dispositivo. Además, el ordenador 164 asociado puede ser local o estar conectado en red con el dispositivo 162 de procesamiento de muestras. En la realización representada, la muestra biológica se puede cargar en el dispositivo 162 de procesamiento de muestras sobre un sustrato 170 de muestras, p. ej., una celda de flujo o portaobjetos, del que se obtienen imágenes para generar datos de secuencia. Por ejemplo, los reactivos que interactúan con la muestra biológica emiten fluorescencia en longitudes de onda particulares en respuesta a un haz de excitación generado por un módulo 172 de obtención de imágenes y, de este modo, devuelven radiación para la obtención de imágenes. Por ejemplo, los componentes fluorescentes pueden generarse mediante ácidos nucleicos marcados con fluorescencia que se hibridan con moléculas complementarias de los componentes o con nucleótidos marcados con fluorescencia que se incorporan en un oligonucleótido usando una polimerasa. Como apreciarán los expertos en la técnica, la longitud de onda a la que se excitan los colorantes de la muestra y la longitud de onda a la que fluorescen dependerán de los espectros de absorción y emisión de los colorantes específicos. Esta radiación devuelta puede propagarse a través de los elementos ópticos de dirección. Este retrohaz puede dirigirse generalmente hacia los elementos ópticos de detección del módulo 172 de imágenes.

Los elementos ópticos de detección del módulo de imágenes pueden basarse en cualquier tecnología adecuada y puede ser, por ejemplo, un sensor de dispositivo de acoplamiento de carga (CCD) que genera datos de imágenes pixeladas basándose en fotones que impactan en ubicaciones en el dispositivo. Sin embargo, se entenderá que también se puede usar cualquiera de diversos otros detectores, incluyendo, aunque no de forma limitativa, una matriz de detectores configurada para operación de retardo e integración (TDI), un detector semiconductor de óxido metálico complementario (CMOS), un detector de fotodiodos de avalanchas (APD), un contador de fotones en modo Geiger o cualquier otro detector adecuado. La detección del modo TDI se puede combinar con la exploración lineal como se describe en la patente estadounidense n.º 7.329.860. Otros detectores útiles se describen, por ejemplo, en las referencias proporcionadas anteriormente en el presente documento en el contexto de diversas metodologías de secuenciación de ácidos nucleicos.

El módulo 172 de captura de imágenes puede estar bajo control del procesador, p. ej., a través de un procesador 174, y el dispositivo 162 receptor de muestras también puede incluir controles 176 de E/S, un bus 78 interno, memoria 180

no volátil, RAM 82 y cualquier otra estructura de memoria, de modo que la memoria sea capaz de almacenar instrucciones ejecutables, y otros componentes de hardware adecuados que pueden ser similares a los descritos con respecto a la FIG. 11. Además, el ordenador 164 asociado también puede incluir un procesador 184, controles 186 de E/S, un módulo 84 de comunicaciones y una arquitectura de memoria que incluye RAM 188 y memoria 190 no volátil, de modo que la arquitectura de memoria sea capaz de almacenar instrucciones 192 ejecutables. Los componentes de hardware pueden estar vinculados mediante un bus 194 interno, que también puede vincularse a la pantalla 196. En realizaciones en las que el dispositivo 160 de secuenciación se implementa como un dispositivo todo en uno, se pueden eliminar determinados elementos de hardware redundantes.

El procesador 184 puede programarse para asignar lecturas de secuenciación individuales a una muestra en función de la secuencia o secuencias índice asociadas según las técnicas proporcionadas en el presente documento. En realizaciones particulares, basándose en los datos de imagen adquiridos por el módulo 172 de imágenes, el dispositivo 160 de secuenciación puede configurarse para generar datos de secuenciación que incluyen identificaciones de nucleótidos para cada base de una lectura de secuenciación. Además, basándose en los datos de la imagen, incluso para lecturas de secuenciación que se realizan en serie, las lecturas individuales pueden vincularse a la misma ubicación a través de los datos de la imagen y, por lo tanto, a la misma cadena molde. De esta manera, las lecturas de secuenciación de índice pueden asociarse con una lectura de secuenciación de una secuencia de inserción antes de asignarse a una muestra de origen. El procesador 184 también puede programarse para realizar análisis posteriores en las secuencias correspondientes a las inserciones para una muestra particular después de la asignación de lecturas de secuenciación a la muestra.

La FIG. 12 es un ejemplo de una pantalla 200 de la interfaz gráfica de usuario que puede ser generada por el dispositivo 160 de secuenciación para la introducción por parte del usuario de información relacionada con reacciones de secuenciación usando los fragmentos de ácido nucleico indexados proporcionados en el presente documento. Por ejemplo, el usuario puede proporcionar información relacionada con un nombre o identificación de cada muestra en la secuenciación, el número de sitios de índice y el conjunto o conjuntos de índices particulares utilizados para cada muestra. En una realización, los conjuntos de índices están disponibles en el mercado y la pantalla 200 de la interfaz de usuario proporciona un menú desplegable de conjuntos de índices disponibles en el mercado. Cada muestra individual puede entonces asociarse con uno o más conjuntos de índices seleccionados disponibles en el mercado. En función de la selección, el procesador (p. ej., el procesador 184) del dispositivo 160 de secuenciación accede a la información de secuencia de índice almacenada correspondiente al conjunto de índices seleccionado desde una memoria y utiliza la información de secuencias índice a la que se accede para asignar lecturas de secuencias índice adquiridas por el dispositivo 160 de secuenciación a muestras particulares. Una vez asignadas a una muestra particular en función de la lectura o lecturas de secuencias índice, las lecturas de secuenciación representativas de la inserción y asociadas con una ubicación de la imagen de la lectura de secuencia índice se asignan conjuntamente a la muestra particular.

Los efectos técnicos de las realizaciones descritas incluyen una indexación mejorada y más precisa de secuencias de ácidos nucleicos. La indexación mejorada puede reducir las lecturas de secuenciación asignadas incorrectamente de una información múltiple (p. ej., de muestras múltiples) a información más significativa para los médicos. Además, las mejoras en la precisión de la asignación de lecturas de secuenciación asociadas con las técnicas actuales facilitan estrategias de secuenciación de alto rendimiento que proporcionan ahorros comerciales y de tiempo. Las secuencias índice proporcionadas en el presente documento abordan el sesgo introducido en los datos de secuenciación que afecta los recuentos de cobertura de secuenciación.

Si bien en el presente documento solo se han ilustrado y descrito determinadas características de la descripción, a los expertos en la técnica se les ocurrirán muchas modificaciones y cambios.

REIVINDICACIONES

1. Un método implementado por ordenador para asignar datos de secuenciación representativos de una secuencia de interés a una muestra, que comprende:

proporcionar datos de secuenciación representativos de una secuencia de interés y proporcionar datos de secuenciación representativos de una secuencia índice 5' y una secuencia índice 3', en donde dichos datos de secuenciación se obtienen a partir de la secuenciación de una pluralidad de fragmentos de ácido nucleico de índice doble generados a partir de una muestra, en donde cada fragmento de ácido nucleico individual de los fragmentos de ácido nucleico de índice doble comprende una secuencia de interés procedente de la muestra, una secuencia adaptadora 5', una secuencia índice 5', una secuencia adaptadora 3' y una secuencia índice 3' para generar fragmentos de ácido nucleico de doble índice, en donde una pluralidad de secuencias índice 5' diferentes seleccionadas de un primer conjunto de secuencias índice 5' asociadas con la muestra y una pluralidad de secuencias índice 3' diferentes seleccionadas de un segundo conjunto de secuencias índice 3' asociadas con la muestra están representadas en los fragmentos de ácido nucleico de índice doble y en donde la pluralidad de secuencias índice 5' diferentes y la pluralidad de secuencias índice 3' diferentes se pueden distinguir entre sí, y procesar dichos datos de secuenciación para asignar una secuencia individual de interés a la muestra solo cuando la secuencia individual de interés está asociada tanto con la secuencia índice 5' seleccionada del primer conjunto como con la secuencia índice 3' seleccionada del segundo conjunto.

2. El método de la reivindicación 1, comprendiendo además el método generar datos de secuencia representativos de la secuencia de interés y generar datos de secuencia representativos de la secuencia índice 5' y la secuencia índice 3'.

3. El método de la reivindicación 1 o 2, en donde la secuencia individual de interés está asociada tanto con la secuencia índice 5' seleccionada del primer conjunto como con la secuencia índice 3' seleccionada del segundo conjunto basándose en la ubicación conjunta de los datos de secuenciación representativos de la secuencia índice 5' y la secuencia índice 3' con los datos de secuenciación representativos de la secuencia de interés.

4. El método de la reivindicación 1, 2 o 3, que comprende asignar datos de secuencia de otra secuencia individual de interés a la muestra en función de la complementariedad con la secuencia individual de interés.

5. El método de una cualquiera de las reivindicaciones 1 a 4, en donde los datos de secuenciación representativos de la secuencia índice 5' y la secuencia índice 3' con los datos de secuenciación representativos de la secuencia de interés se generan a partir de una misma cadena individual de los fragmentos de ácido nucleico de índice doble.

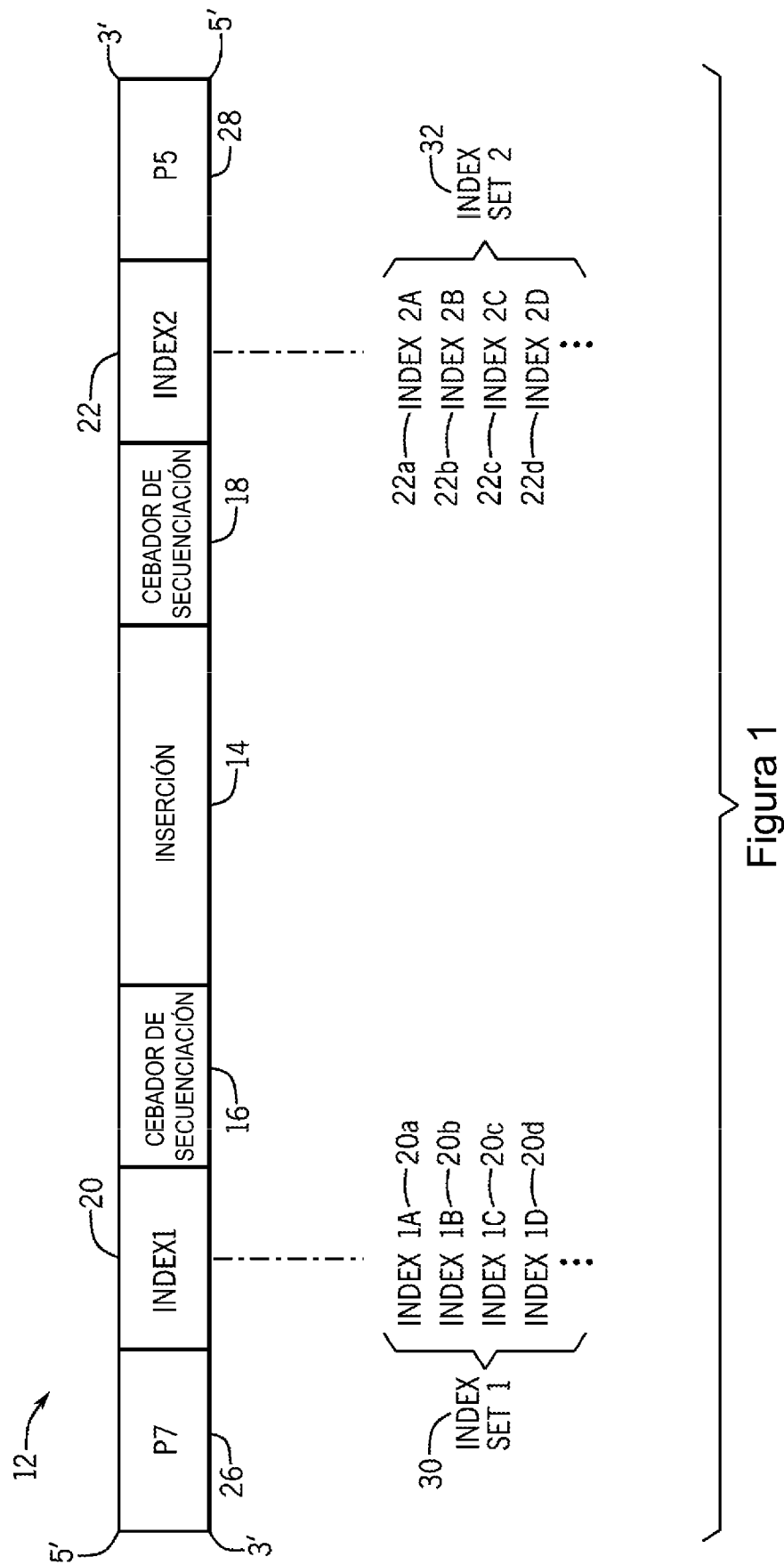
6. El método de una cualquiera de las reivindicaciones 1 a 5, en donde los datos de secuenciación representativos de la secuencia índice 5' y la secuencia índice 3' se generan a partir de diferentes cadenas individuales de los fragmentos de ácido nucleico de índice doble.

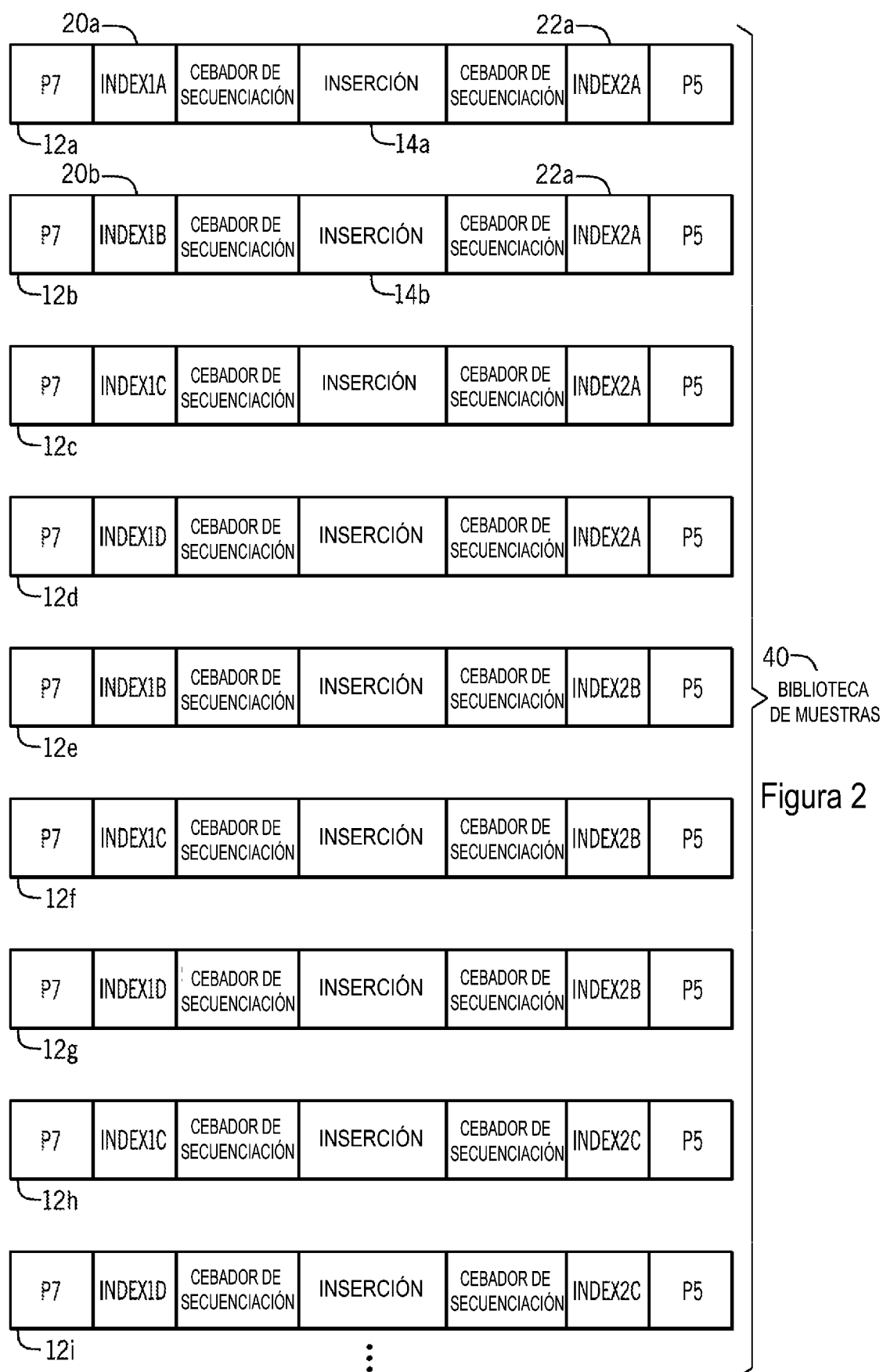
7. El método de la reivindicación 1, que comprende poner en contacto los fragmentos de ácido nucleico de índice doble con un sustrato de secuenciación que comprende moléculas de captura inmovilizadas, las moléculas de captura configuradas para hibridar con las secuencias adaptadora 5' y adaptadora 3' de los fragmentos de ácido nucleico.

8. El método de la reivindicación 1, que comprende eliminar secuencias individuales en los datos de secuenciación que incluyen solo una de la secuencia índice 5' o la secuencia índice 3'.

9. El método de la reivindicación 1 u 8, que comprende además proporcionar fragmentos de ácido nucleico de doble índice de una segunda muestra, en donde los fragmentos de ácido nucleico de doble índice de la segunda muestra se modifican con la secuencia adaptadora 5' y la secuencia adaptadora 3' y en donde una segunda pluralidad de secuencias índice 5' diferentes y una segunda pluralidad de secuencias índice 3' diferentes están representadas en los fragmentos de ácido nucleico de doble índice de la segunda muestra y en donde la segunda pluralidad de secuencias índice 5' diferentes y una segunda pluralidad de secuencias índice 3' diferentes se pueden distinguir entre sí y de la pluralidad de secuencias índice 5' diferentes y la pluralidad de secuencias índice 3' diferentes.

10. El método de la reivindicación 9, que comprende poner en contacto simultáneamente los fragmentos de ácido nucleico de índice doble de la muestra y la segunda muestra con un sustrato de secuenciación que comprende moléculas de captura inmovilizadas, las moléculas de captura configuradas para hibridar con las secuencias adaptadora 5' y adaptadora 3' de los fragmentos de ácido nucleico.





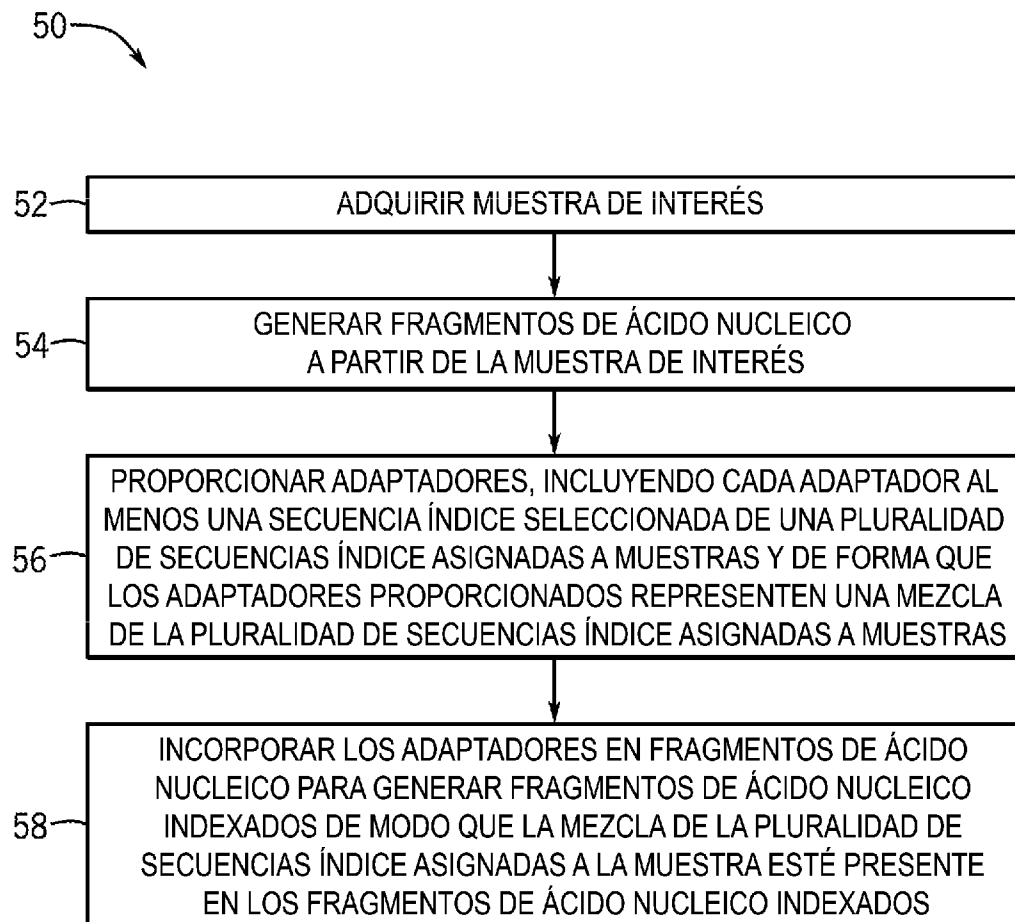
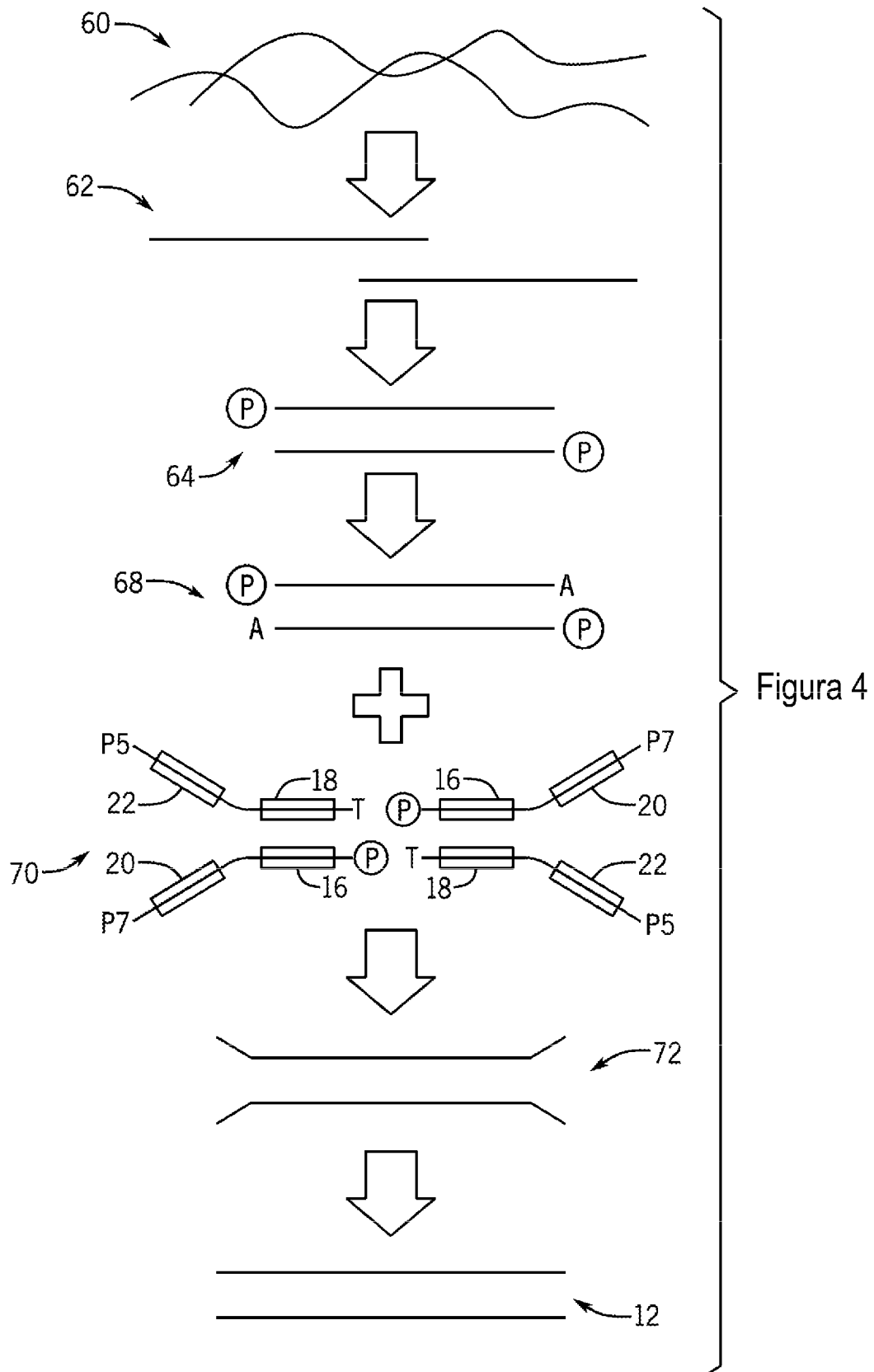


Figura 3



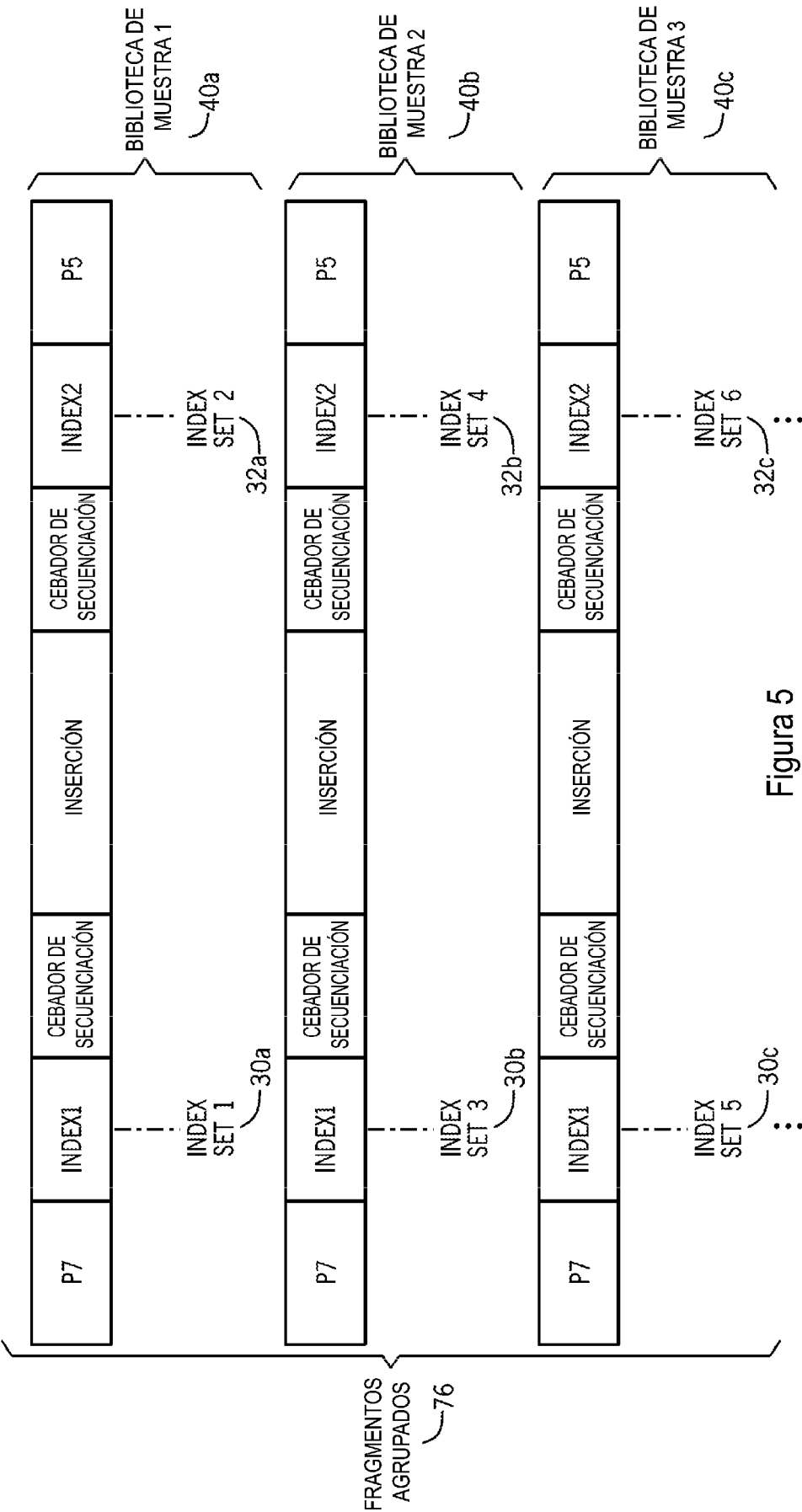


Figura 5

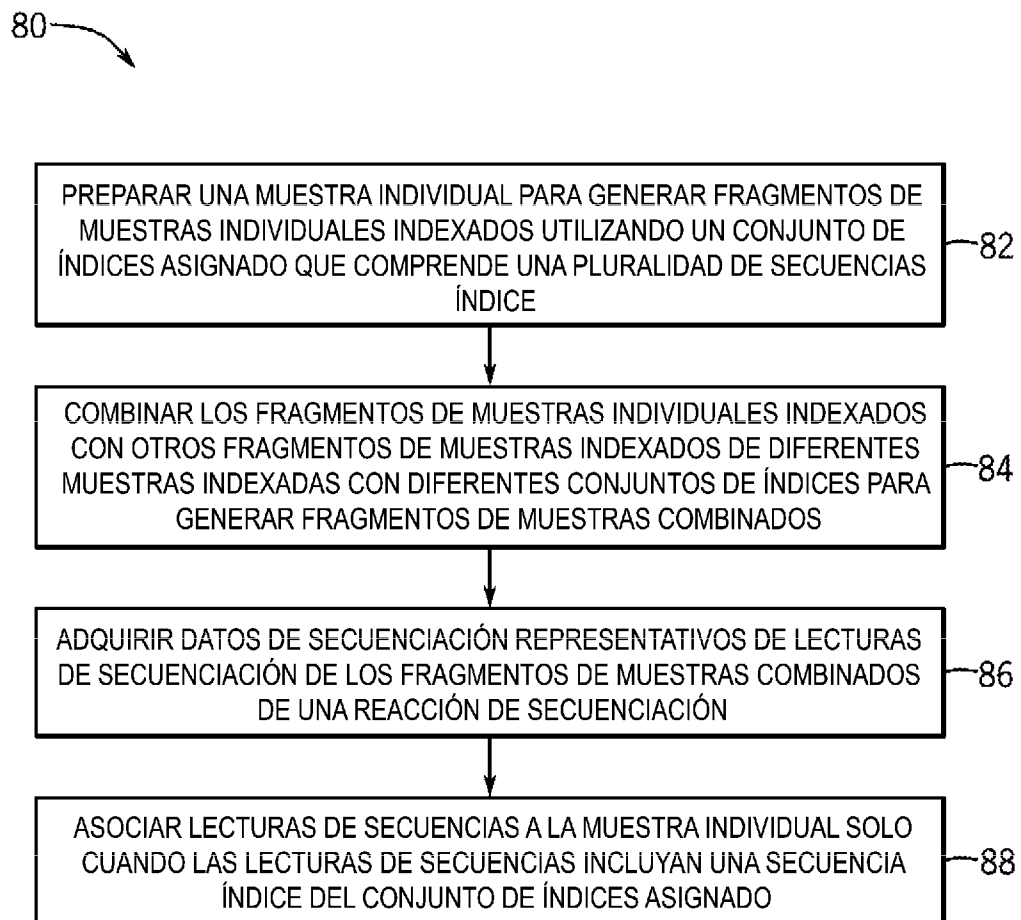
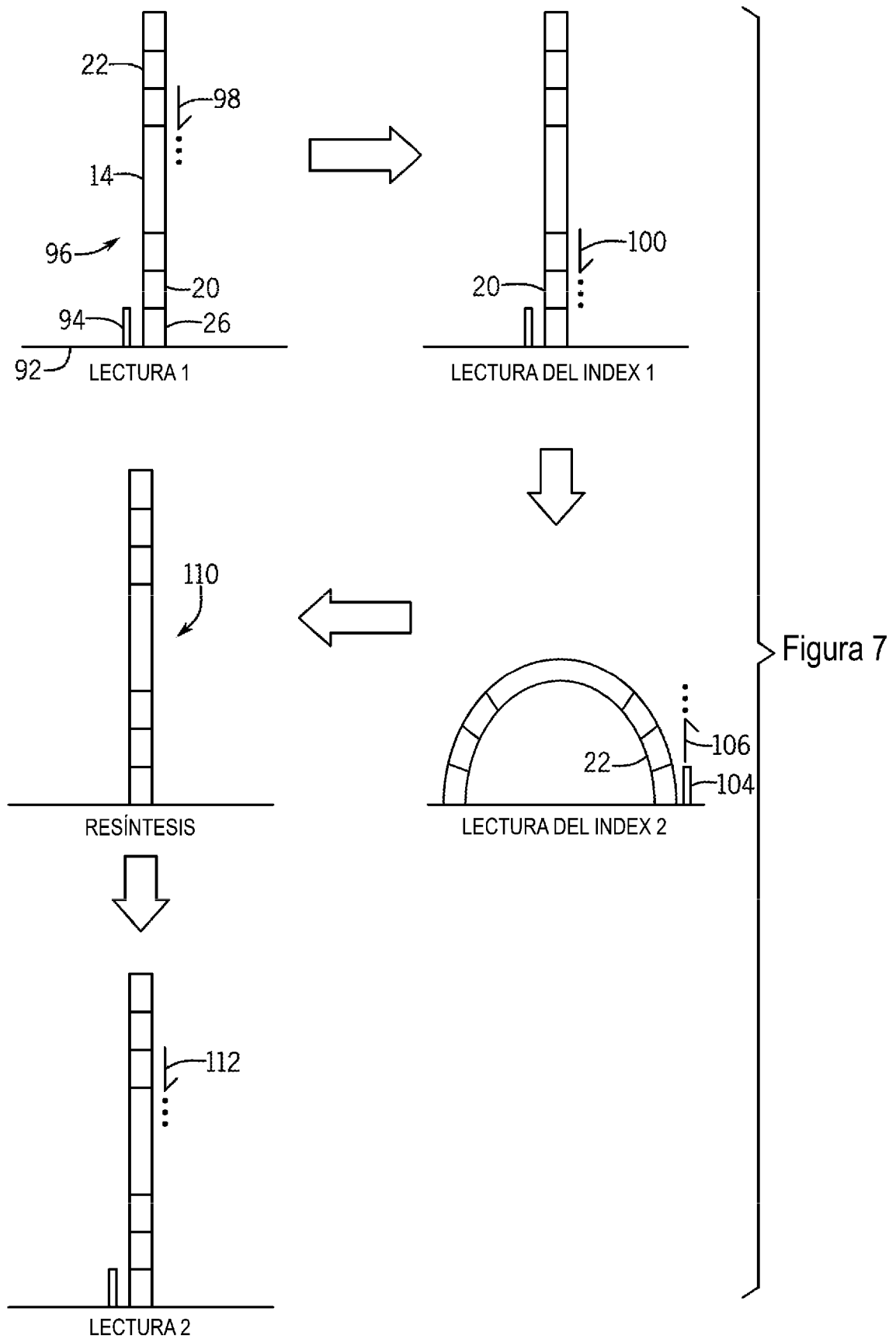


Figura 6



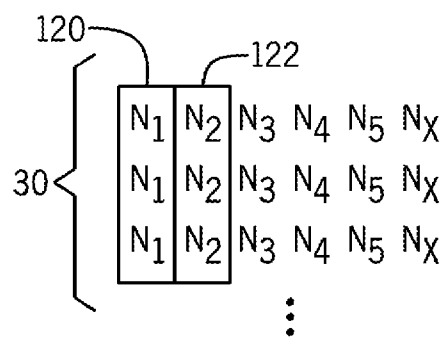


Figura 8

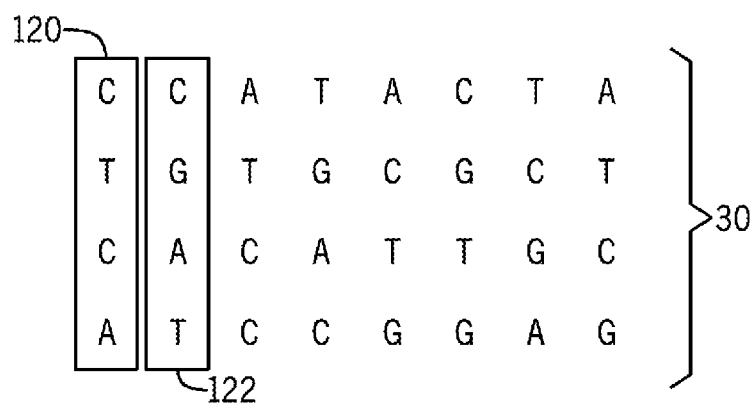


Figura 9

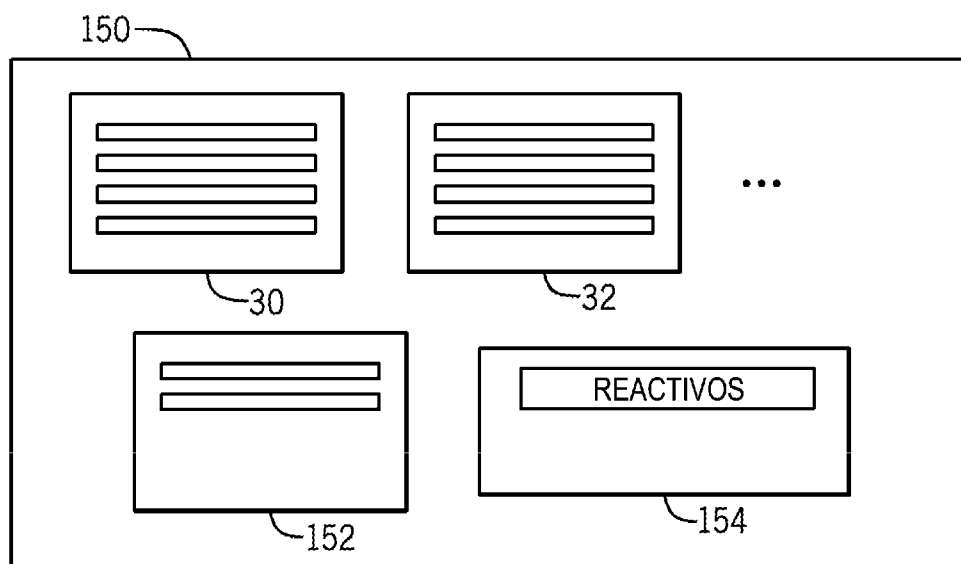


Figura 10

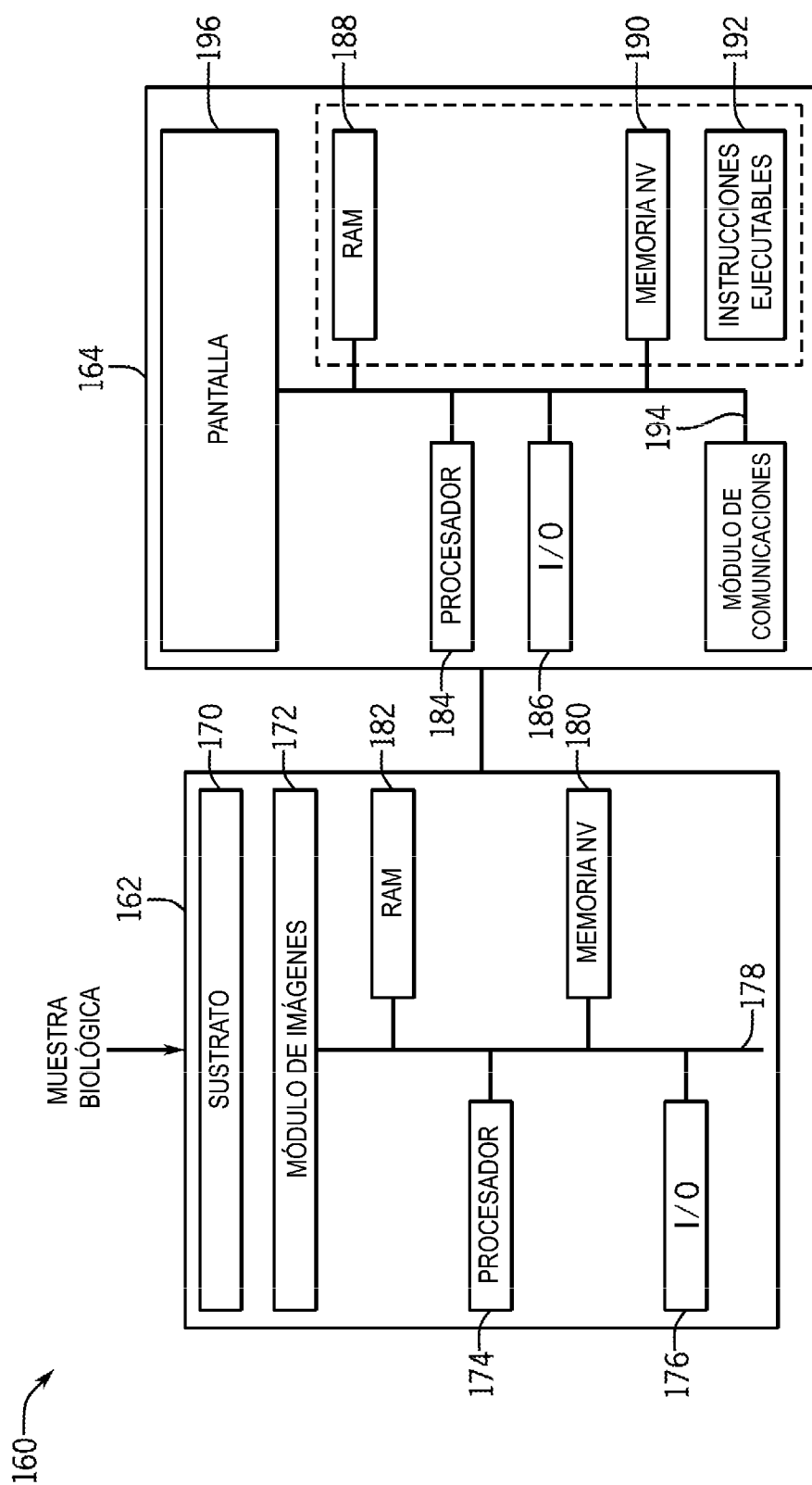


Figura 11

200

/ (1) (1) GESTOR DE CICLO LOCAL x
 - □ ×

← → ↻ 🏠 📄
 ≡

APLICACIONES

GESTOR DE CICLO LOCAL
 🔍 ⚙ ADMINISTRADOR USUARIO ▾

CREAR CICLO

NOMBRE DEL CICLO*
 DESCRIPCIÓN DEL CICLO

CONFIGURACIÓN DEL CICLO

KIT DE PREPARACIÓN PARA LA BIBLIOTECA*
 TIPO DE LECTURA*

LECTURAS DE INDEX* ☐ 0 ☐ 1 ☒ 2

☒ LECTURA ÚNICA ☐ LECTURA DE INDEX ☐ INDEX EMPAREJADO ☐ LECTURA FINAL

	1	1	2	2
DURACIÓN DE LA LECTURA*	51	6	8	...

CEBADORES PERSONALIZADOS ☐ LECTURA 1 ☐ INDEX 1 ☐ INDEX 2 ...

AJUSTES ESPECÍFICOS DEL MÓDULO

❗ NO HAY CONFIGURACIÓN ESPECÍFICA DEL MÓDULO DISPONIBLE PARA EL KIT DE PREPARACIÓN DE BIBLIOTECA SELECCIONADO.

MOSTRAR CONFIGURACIÓN AVANZADA DEL MÓDULO

IMPORTAR MUESTRAS
IMPORTAR MANIFIESTO

PLANTILLA
EXPORTAR

	ID DE MUESTRA*	NOMBRE DE LA MUESTRA*	DESCRIPCIÓN DE LA MUESTRA	INDEX 1 (17)*	INDEX 2 (15)*	MANIFIESTO*	NORMALIZACIÓN GENÉTICA	
1								×
2								×

+ 1 FILAS

CANCELAR
GUARDAR CICLO

Figura 12