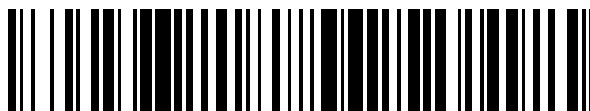


19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 808 824**

51 Int. Cl.:

C12Q 1/6869 (2008.01)

G16B 30/00 (2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **03.12.2014** **PCT/US2014/068409**

87 Fecha y número de publicación internacional: **11.06.2015** **WO15084985**

96 Fecha de presentación y número de la solicitud europea: **03.12.2014** **E 14867596 (0)**

97 Fecha y número de publicación de la concesión europea: **03.06.2020** **EP 3077943**

54 Título: **Métodos y sistemas para analizar datos de imagen**

30 Prioridad:

03.12.2013 US 201361911319 P

12.12.2013 US 201361915426 P

12.12.2013 US 201361915455 P

45 Fecha de publicación y mención en BOPI de la traducción de la patente:
02.03.2021

73 Titular/es:

ILLUMINA, INC. (100.0%)

5200 Illumina Way

San Diego, CA 92122, US

72 Inventor/es:

BELITZ, PAUL;

TANNER, STEPHEN;

VIECELI, JOHN, S. y

CHEN, XIAOYU

74 Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

ES 2 808 824 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

Métodos y sistemas para analizar datos de imagen

5 Referencia cruzada a solicitudes relacionadas

La presente solicitud reivindica prioridad a y el beneficio de las solicitudes provisionales de Estados Unidos con Números 61/911.319, presentada el 3 de diciembre de 2013; 61/915.455, presentada el 12 de diciembre de 2013; y 61/915.426, presentada el 12 de diciembre de 2013.

10 Antecedentes

El análisis de datos de imagen presenta un número de retos, especialmente con respecto a la comparación de imágenes de un artículo o estructura que se capturan desde diferentes puntos de referencia. Un campo que ilustra muchos de estos retos es el de análisis de secuencias de ácido nucleico.

La detección de secuencias de ácido nucleico específicas presentes en una muestra biológica tiene una amplia variedad de aplicaciones, tal como identificar y clasificar microorganismos, diagnosticar enfermedades infecciosas, detectar o caracterizar anomalías genéticas, identificar cambios genéticos asociados con el cáncer, estudiar la susceptibilidad genética a enfermedades y medir la respuesta a varios tipos de tratamiento. Una técnica valiosa de detección de secuencias de ácido nucleico específicas en una muestra biológica es la secuenciación de ácido nucleico.

La metodología de secuenciación de ácido nucleico ha evolucionado significativamente a partir de los métodos de degradación química usados por Maxam y Gilbert y los métodos de elongación de cadena usados por Sanger. Actualmente, existen diversos procesos diferentes que se emplean para dilucidar una secuencia de ácido nucleico. Un proceso de secuenciación particularmente popular es la secuenciación por síntesis. Una razón de su popularidad es que ésta técnica puede aplicarse fácilmente a proyectos de secuenciación masivamente paralelos. Por ejemplo, usando una plataforma automatizada, es posible efectuar cientos de miles de reacciones de secuenciaciones simultáneamente. La secuenciación por síntesis difiere del clásico enfoque de secuenciación de dideoxi en el que, en lugar de generar un gran número de secuencias y a continuación caracterizar las mismas en una etapa posterior, se emplea una supervisión en tiempo real de la incorporación de cada base en una cadena creciente. Aunque este enfoque podría verse como lento en el contexto de una reacción de secuenciación individual, puede usarse para generar grandes cantidades de información de secuencia en cada ciclo de secuenciación cuando se realizan en paralelo cientos de miles a millones de reacciones. A pesar de estas ventajas, el gran tamaño y cantidad de información de secuencias obtenida a través de tales métodos pueden limitar la velocidad y calidad de análisis de datos de secuencias. Por lo tanto, existe una necesidad de métodos y sistemas que mejoren la velocidad y precisión del análisis de datos de secuenciación de ácido nucleico.

Kirchner et al., "Improved base Calling for the Illumina Genome Analyzer using machine learning strategies" GENOME BIOLOGY, BIOMED CENTRAL LTD., LONDRES, REINO UNIDO, vol. 10, n.º 8, describe un método y sistema de identificación de un nucleótido a partir de datos de secuenciación de ácido nucleico e identificación de bases de cuatro canales. Se obtienen cuatro imágenes y se analizan cuatro valores de intensidad. Se aplica una corrección de ajuste de fase a los datos de secuenciación para obtener valores de intensidad corregidos.

45 Breve resumen

En este documento se proporcionan métodos para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación. En algunas realizaciones, los métodos pueden comprender las etapas de: calcular un conjunto de valores de predictor para la identificación de bases; y, a continuación, usando los valores de predictor para buscar una puntuación de calidad en una tabla de calidad. En algunas realizaciones, la lectura de secuenciación utiliza identificación de bases de dos canales. En algunas realizaciones, la lectura de secuenciación utiliza identificación de bases de un canal. En ciertos aspectos, la tabla de calidad se genera usando puntuación Phred en un conjunto de datos de calibración, siendo el conjunto de calibración representativo de variabilidad de ejecución y secuencia. En ciertos aspectos, los valores de predictor se seleccionan a partir del grupo que consiste en: solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; y solapamiento de señal con fondo (SOWB). En ciertos aspectos, el conjunto de valores de predictor comprende solapamiento en línea; pureza; ajuste de fase; y start5. En ciertos aspectos, el conjunto de valores de predictor comprende puntuación de hexámeros; y acumulación de motivos.

En ciertos aspectos, el método comprende adicionalmente las etapas de: descontar puntuaciones de calidad no fiables al final de cada lectura; identificar lecturas en las que la segunda peor depuración en las primeras 25 identificaciones de bases está por debajo de un umbral preestablecido; y marcar las lecturas como datos de baja calidad. En ciertos aspectos, el método comprende adicionalmente usar un algoritmo para identificar un umbral de fiabilidad. En ciertos aspectos, identificaciones de bases fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por encima del umbral e identificaciones de bases no fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por debajo del umbral. En ciertos aspectos, el

algoritmo comprende un algoritmo de Segmentos de Puntuación Máxima de Extremo Anclado (EAMSS). En ciertos aspectos, el algoritmo usa un Modelo Oculto de Markov que identifica desplazamientos en las distribuciones locales de puntuaciones de calidad.

- 5 También se proporciona en este documento un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación, comprendiendo el sistema: un procesador; una capacidad de almacenamiento; y un programa para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación, comprendiendo el programa instrucciones para: calcular un conjunto de valores de predictor para la identificación de bases; y, a continuación, usando los valores de predictor para buscar una puntuación de calidad en una tabla de
10 calidad. En ciertos aspectos, la tabla de calidad se genera usando puntuación Phred en un conjunto de datos de calibración, siendo el conjunto de calibración representativo de variabilidad de ejecución y secuencia. En ciertos aspectos, los valores de predictor se seleccionan a partir del grupo que consiste en: solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; y solapamiento de señal con fondo (SOWB). En ciertos aspectos,
15 el conjunto de valores de predictor comprende solapamiento en línea; pureza; ajuste de fase; y start5. En ciertos aspectos, el conjunto de valores de predictor comprende puntuación de hexámeros; y acumulación de motivos.

- En ciertos aspectos, el sistema puede comprender adicionalmente instrucciones para: descontar puntuaciones de calidad no fiables al final de cada lectura; identificar lecturas en las que la segunda peor depuración en las primeras
20 25 identificaciones de bases está por debajo de un umbral preestablecido; y marcar las lecturas como datos de baja calidad. En ciertos aspectos, el sistema comprende además instrucciones para usar un algoritmo para identificar un umbral de fiabilidad. En ciertos aspectos, las identificaciones de bases fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por encima del umbral e identificaciones de bases no fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por debajo del umbral.
25 En ciertos aspectos, el algoritmo comprende un algoritmo de Segmentos de Puntuación Máxima de Extremo Anclado (EAMSS). En ciertos aspectos, el algoritmo usa un Modelo Oculto de Markov que identifica desplazamientos en las distribuciones locales de puntuaciones de calidad.

- También se presentan en este documento métodos y sistema para generar un valor de intensidad con corrección de
30 ajuste de fase. Los métodos pueden comprender: realizar una pluralidad de ciclos de una secuenciación por reacción de síntesis de tal forma que, en cada ciclo, se genera una señal indicativa de incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con lo que una porción de la señal es ruido asociado con un nucleótido incorporado durante un ciclo anterior; detectar la señal en cada ciclo, teniendo la señal un valor de intensidad; y corregir el valor de intensidad para ajuste de fase aplicando una corrección de ajuste de fase de primer orden al valor de
35 intensidad; en el que se calcula una nueva corrección de ajuste de fase de primer orden para cada ciclo.

- En algunos aspectos, la corrección de ajuste de fase de primer orden comprende restar un valor de intensidad del ciclo inmediatamente anterior del valor de intensidad del ciclo actual. El método puede comprender adicionalmente restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual. En algunos
40 aspectos, la corrección de ajuste de fase comprende: $I_{(\text{ciclo})\text{corregido}} = I_{(\text{ciclo})\text{N}} - X * I_{(\text{ciclo})\text{N}-1} - Y * I_{(\text{ciclo})\text{N}+1}$. En ciertos aspectos, los valores de X y/o Y se eligen para optimizar una determinación de depuración. En ciertos aspectos, la determinación de depuración comprende depuración media. En ciertos aspectos, la ejecución de secuenciación puede utilizar identificación de bases de un canal, dos canales o cuatro canales.

- También se presentan en este documento sistemas para generar un valor de intensidad con corrección de ajuste de
45 fase. Los sistemas pueden comprender: un procesador; una capacidad de almacenamiento; y un programa para generar un valor de intensidad con corrección de ajuste de fase, comprendiendo el programa instrucciones para: realizar una pluralidad de ciclos de una secuenciación por reacción de síntesis de tal forma que, en cada ciclo, se genera una señal indicativa de incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con
50 lo que una porción de la señal es ruido asociado con un nucleótido incorporado durante un ciclo anterior; detectar la señal en cada ciclo, teniendo la señal un valor de intensidad; y corregir el valor de intensidad para ajuste de fase aplicando una corrección de ajuste de fase de primer orden al valor de intensidad; en el que se calcula una nueva corrección de ajuste de fase de primer orden para cada ciclo.

- En algunos aspectos, la corrección de ajuste de fase de primer orden comprende restar un valor de intensidad del ciclo inmediatamente anterior del valor de intensidad del ciclo actual. El método puede comprender adicionalmente restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual. En algunos
55 aspectos, la corrección de ajuste de fase comprende: $I_{(\text{ciclo})\text{corregido}} = I_{(\text{ciclo})\text{N}} - X * I_{(\text{ciclo})\text{N}-1} - Y * I_{(\text{ciclo})\text{N}+1}$. En ciertos aspectos, los valores de X y/o Y se eligen para optimizar una determinación de depuración. En ciertos aspectos, la determinación de depuración comprende depuración media. En ciertos aspectos, la ejecución de secuenciación puede utilizar
60 identificación de bases de un canal, dos canales o cuatro canales.

- También se presentan en este documento métodos y sistemas para identificar una base de nucleótido a partir de datos
65 de secuenciación en los que se obtienen dos imágenes separadas de una matriz de características en una superficie. En algunas realizaciones, el método comprende: detectar la presencia o ausencia de una señal en dos canales diferentes para cada una de una pluralidad de características en una matriz en un momento particular, generando de

este modo un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una de las características, en el que la combinación de valores de intensidad en cada uno de los dos canales corresponde a una de cuatro diferentes bases de nucleótidos; ajustar cuatro distribuciones gaussianas a los valores de intensidad, teniendo cada distribución un centroide; calcular un valor de probabilidad que indica la probabilidad de una característica particular que pertenece a cada una de las cuatro distribuciones; y seleccionar para cada característica de dicha pluralidad de características la distribución que tiene el mayor valor de probabilidad, en el que dicha distribución corresponde a la identidad de la base de nucleótido presente en dicha característica particular.

También se presenta en este documento un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación, comprendiendo el sistema: un procesador; una capacidad de almacenamiento; y un programa para identificar una base de nucleótido, comprendiendo el programa instrucciones para: detectar la presencia o ausencia de una señal en dos canales diferentes para cada una de una pluralidad de características en una matriz en un momento particular, generando de este modo un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una de las características, en el que la combinación de valores de intensidad en cada uno de los dos canales corresponde a una de cuatro diferentes bases de nucleótidos; ajustar cuatro distribuciones gaussianas a los valores de intensidad, teniendo cada distribución un centroide; calcular un valor de probabilidad que indica la probabilidad de una característica particular que pertenece a cada una de las cuatro distribuciones; y seleccionar para cada característica de dicha pluralidad de características la distribución que tiene el mayor valor de probabilidad, en el que dicha distribución corresponde a la identidad de la base de nucleótido presente en dicha característica particular.

También se presenta en este documento un método de identificación de una base de nucleótido, comprendiendo el método: obtener un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una pluralidad de características en una matriz, en el que el valor de intensidad para cada característica en uno o ambos conjuntos corresponde a la presencia o ausencia de una base de nucleótido particular de entre cuatro posibles bases de nucleótidos en la característica; ajustar cuatro distribuciones gaussianas a los valores de intensidad, teniendo cada distribución un centroide; calcular cuatro valores de probabilidad para cada característica, en el que cada valor de probabilidad indica la probabilidad de una característica particular que pertenece a una de las cuatro distribuciones; y seleccionar para cada característica de dicha pluralidad de características la distribución que tiene el mayor de los cuatro valores de probabilidad, en el que la distribución corresponde a la identidad de la base de nucleótido presente en la característica particular.

También se presenta en este documento un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación, comprendiendo el sistema: un procesador; una capacidad de almacenamiento; y un programa para identificar una base de nucleótido, comprendiendo el programa instrucciones para: obtener un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una pluralidad de características en una matriz, en el que el valor de intensidad para cada característica en uno o ambos conjuntos corresponde a la presencia o ausencia de una base de nucleótido particular de entre cuatro posibles bases de nucleótidos en la característica; ajustar cuatro distribuciones gaussianas a los valores de intensidad, teniendo cada distribución un centroide; calcular cuatro valores de probabilidad para cada característica, en el que cada valor de probabilidad indica la probabilidad de una característica particular que pertenece a una de las cuatro distribuciones; y seleccionar para cada característica de dicha pluralidad de características la distribución que tiene el mayor de los cuatro valores de probabilidad, en el que la distribución corresponde a la identidad de la base de nucleótido presente en la característica particular.

En cualquiera de los métodos y sistemas descritos anteriormente, ciertos aspectos pueden incluir realizaciones en las que ajustar puede comprender usar uno o más algoritmos a partir del grupo que consiste en: un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, maximización de la expectativa y método basado en histogramas. En algunos aspectos, ajustar puede comprender usar un algoritmo de maximización de la expectativa. En algunos aspectos, el método puede comprender normalizar los valores de intensidad. En ciertos aspectos, se calcula un valor de depuración para cada característica. En ciertos aspectos, el valor de depuración es una función de la distancia relativa desde una característica a los dos centroides más cercanos gaussianos. En algunos aspectos, se filtran características que tienen un valor de depuración por debajo de un valor umbral.

Los detalles de una o más realizaciones se exponen en los dibujos adjuntos y la descripción a continuación. Otras características, objetos y ventajas serán evidentes a partir de la descripción y los dibujos, y a partir de las reivindicaciones.

Breve descripción de los dibujos

Las Figuras 1A y 1B representan datos de intensidad para un sistema de dos canales. La Figura 1A es un gráfico de dispersión que muestra intensidades sin tratar para una placa particular y un ciclo particular, en el que el nucleótido C se representa mediante señal en canal 1 únicamente, el nucleótido A se representa mediante señal en canal 2 únicamente, el nucleótido T se representa mediante señal en ambos canales 1 y 2, y el nucleótido G está "oscuro". La Figura 1B muestra intensidades con corrección de ajuste de fase de los mismos datos usando una corrección de ajuste de fase de acuerdo con una realización de los métodos presentados en este documento.

La Figura 2 representa datos de intensidad para un sistema de dos canales que se ha sometido a diversas correcciones de ajuste de fase.

La Figura 3 muestra un gráfico ilustrativo de intensidades de imagen a partir de secuenciación de dos canales.

La Figura 4 muestra un enfoque para ajustar distribuciones gaussianas a datos de intensidad de dos canales, de acuerdo con una realización.

La Figura 5 expone una aplicación de maximización de la expectativa a datos de secuenciación de un canal (imagen izquierda) y datos de secuenciación de dos canales (imagen derecha).

La Figura 6 es un diagrama de flujo que ilustra un método de acuerdo con una realización.

La Figura 7 es un diagrama de flujo que ilustra un método de acuerdo con una realización.

La Figura 8 es un diagrama de flujo que ilustra un método de acuerdo con una realización.

La Figura 9 es un diagrama de flujo que ilustra un método de acuerdo con una realización.

La Figura 10 es un diagrama de bloques de un sistema de acuerdo con una realización.

Descripción detallada

La presente solicitud describe diversos métodos y sistemas para llevar a cabo los métodos. Ejemplos de algunos de los métodos se describen como una serie de etapas. Sin embargo, debería entenderse que las realizaciones no se limitan a las etapas particulares y/u orden de las etapas descritas en este documento. Pueden omitirse etapas, pueden modificarse etapas y/o pueden añadirse otras etapas. Además, las etapas descritas en este documento pueden combinarse, etapas pueden realizarse simultáneamente, etapas pueden realizarse concurrentemente, etapas pueden dividirse en múltiples subetapas, etapas pueden realizarse en un orden diferente o etapas (o una serie de etapas) pueden realizarse de nuevo de una forma iterativa. Además, aunque se exponen diferentes métodos en este documento, debería entenderse que los diferentes métodos (o etapas de los diferentes métodos) pueden combinarse en otras realizaciones.

El análisis de datos de imagen presenta un número de retos, especialmente con respecto a la comparación de imágenes de un artículo o estructura que se capturan desde diferentes puntos de referencia. La mayoría de metodología de análisis de imágenes emplea, al menos en parte, etapas para alinear múltiples imágenes separadas entre sí basándose en características o elementos presentes en ambas imágenes. Diversas realizaciones de las composiciones y métodos divulgados en este documento mejoran métodos anteriores para el análisis de imágenes. Algunos métodos anteriores para análisis de imágenes se exponen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537 presentada en 13 de enero de 2011 y titulada, "DATA PROCESSING SYSTEM AND METHODS". Realizaciones descritas en lo sucesivo también se describen en la Solicitud Provisional de Estados Unidos N.º 61/911.319, presentada el 3 de diciembre de 2013. También pueden usarse una o más realizaciones con realizaciones descritas en la Publicación de Solicitud de Patente de Estados Unidos N.º 2016/0085910, publicada el 24 de marzo de 2016.

En la actualidad, se han desarrollado herramientas que adquieren y analizan datos de imagen generados en puntos de tiempo o perspectivas diferentes. Algunos ejemplos incluyen herramientas para análisis de imágenes de satélite y herramientas de biología molecular para secuenciar y caracterizar la identidad molecular de un espécimen. En cualquier tal sistema, adquirir y almacenar grandes números de imágenes de alta calidad habitualmente requiere cantidades masivas de capacidad de almacenamiento. Adicionalmente, una vez adquiridos y almacenados, el análisis de datos de imagen puede volverse intensivo en recursos y puede interferir con capacidad de procesamiento de otras funciones, tales como adquisición constante y almacenamiento de datos de imagen adicionales. Como tal, serían beneficiosos métodos y sistemas que mejoren la velocidad y precisión de análisis de la adquisición y análisis de datos de imagen.

En el campo de la biología molecular, uno de los procesos para secuenciación de ácido nucleico en uso es secuenciación por síntesis. La técnica puede aplicarse a proyectos de secuenciación masivamente paralelos. Por ejemplo, usando una plataforma automatizada, es posible efectuar cientos de miles de reacciones de secuenciaciones simultáneamente. Por lo tanto, la divulgación se refiere a instrumentos y métodos para adquirir, almacenar y analizar datos de imagen generados durante secuenciación de ácido nucleico.

Enormes ganancias en la cantidad de datos que pueden adquirirse y almacenarse hacen incluso más beneficiosos a los métodos de análisis de imagen simplificados. Por ejemplo, los métodos de análisis de imagen descritos en este documento permiten que tanto diseñadores como usuarios finales hagan un uso eficiente de hardware informático existente. Por consiguiente, en este documento se presentan métodos y sistemas que reducen la carga de cálculo del procesamiento de datos ante el rápido crecimiento de la emisión de datos. Por ejemplo, en el campo de secuenciación de ADN, los rendimientos han escalado 15 veces durante el transcurso del último año, y ahora puede alcanzarse cientos de gigabases en una única ejecución de un dispositivo de secuenciación de ADN. Si los requisitos de infraestructura de cálculo crecen exponencialmente, los experimentos de genoma a gran escala permanecerían fuera del alcance de la mayoría de investigadores. Por lo tanto, la generación de más datos de secuencia sin procesar aumentará la necesidad de análisis secundario y almacenamiento de datos, haciendo extremadamente valiosa la optimización de transporte y almacenamiento de datos. Algunas realizaciones de los métodos y sistemas presentados en este documento pueden reducir el tiempo, hardware, interconexión de redes y requisitos de infraestructura de laboratorio necesarios para producir datos de secuencia usables.

Como se usa en este documento, una "característica" es un área de interés dentro de un espécimen o campo de visión. Cuando se usa en conexión con dispositivos de micromatriz u otros dispositivos de analítica molecular, una característica se refiere al área ocupada por moléculas similares o idénticas. Por ejemplo, una característica puede ser un oligonucleótido amplificado o cualquier otro grupo de un polinucleótido o polipéptido con una misma o similar

- 5 secuencia. En otras realizaciones, una característica puede ser cualquier elemento o grupo de elementos que ocupa un área física en un espécimen. Por ejemplo, una característica podría ser una parcela de tierra, una masa de agua o similar. Cuando se representa una característica, cada característica tendrá alguna área. Por lo tanto, en muchas realizaciones, una característica no es meramente un píxel.
- 10 Las distancias entre características pueden describirse de cualquier número de formas. En algunas realizaciones, las distancias entre características pueden describirse a partir del centro de una característica hasta el centro de otra característica. En otras realizaciones, las distancias pueden describirse desde el borde de una característica hasta el borde de otra característica, o entre los puntos identificables más externos de cada característica. El borde de una característica puede describirse como el límite físico teórico o real en un chip, o algún punto dentro del límite de la
- 15 característica. En otras realizaciones, las distancias pueden describirse en relación con un punto fijo en el espécimen o en la imagen del espécimen.

- Pueden secuenciarse múltiples copias de ácidos nucleicos en una característica, por ejemplo, proporcionando una base de nucleótido etiquetada a la matriz de moléculas, extendiendo de este modo un cebador hibridado a un ácido nucleico dentro de una característica para producir una señal que corresponde a una característica que comprende el ácido nucleico. En realizaciones preferidas, los ácidos nucleicos dentro de una característica son idénticos o sustancialmente idénticos entre sí.
- 20

- En algunos de los métodos de análisis de imagen descritos en este documento, cada imagen en el conjunto de imágenes incluye señales de colores, en el que un color diferente corresponde a una base de nucleótido diferente. En algunos aspectos, cada imagen del conjunto de imágenes comprende señales que tienen un único color seleccionado de al menos cuatro colores diferentes. En ciertos aspectos, cada imagen en el conjunto de imágenes comprende señales que tienen un único color seleccionado de cuatro colores diferentes.
- 25

- Con respecto a ciertos métodos de cuatro canales descritos en este documento, pueden secuenciarse ácidos nucleicos proporcionando, cuatro diferentes bases de nucleótidos etiquetadas a la matriz de moléculas para producir cuatro imágenes diferentes, comprendiendo cada imagen señales que tienen un único color, en el que el color de señal es diferente para cada una de las cuatro imágenes diferentes, produciendo de este modo un ciclo de cuatro imágenes de color que corresponde a los cuatro posibles nucleótidos presentes en una posición particular en el ácido nucleico. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de color.
- 30
- 35

- Con respecto a ciertos métodos de dos canales descritos en este documento, pueden secuenciarse ácidos nucleicos utilizando métodos y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, puede secuenciarse un ácido nucleico proporcionando un primer tipo de nucleótido que se detecta en un primer canal, un segundo tipo de nucleótido que se detecta en un segundo canal, un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o está mínimamente, en ningún canal. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de color.
- 40
- 45

Puntuación de calidad

- La puntuación de calidad se refiere al proceso de asignar una puntuación de calidad a cada identificación de bases. En algunas realizaciones en las que se detectan cuatro diferentes nucleótidos usando menos de cuatro etiquetas diferentes, la identificación de bases requiere un conjunto diferente de enfoques analíticos en comparación con sistemas que usan detección tradicional de cuatro etiquetas. Como un ejemplo, puede realizarse SBS utilizando métodos de dos canales y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Por ejemplo, en las realizaciones que hacen uso de detección de dos canales, se realiza identificación de bases extrayendo datos de imagen de dos imágenes, en lugar de cuatro. Debido a las diferencias fundamentales implicadas en identificación de bases de dos canales, enfoques tradicionales puntuación de calidad según se aplican a identificación de bases de cuatro canales no son compatibles con datos de identificación de bases de dos canales. Por ejemplo, el perfil de errores presentado por datos de dos canales es fundamentalmente diferente del perfil de errores de datos de cuatro canales. En vista de estas diferencias, se requiere un nuevo enfoque para evaluar la calidad de una identificación de bases.
- 50
- 55
- 60

- Por consiguiente, en este documento se presentan métodos y sistemas para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación. En algunas realizaciones, la lectura de secuenciación utiliza identificación de bases de dos canales. En algunas realizaciones, la lectura de secuenciación utiliza identificación de bases de un canal.
- 65

La puntuación de calidad se cita habitualmente como QXX donde XX es la puntuación y significa que esa identificación particular tiene una probabilidad de error de $10^{-(XX/10)}$. Por ejemplo Q30 es igual a una tasa de error de 1 de 1000, o 0,1 % y Q40 es igual a una tasa de error de 1 de 10.000 o 0,01 %.

En algunas realizaciones, se genera una tabla de calidad usando puntuación Phred en un conjunto de datos de calibración, siendo el conjunto de calibración representativo de variabilidad de ejecución y secuencia. La puntuación Phred se describe en mayor detalle en la Patente de Estados Unidos N.º 8.392.126 titulada, "METHOD AND SYSTEM FOR DETERMINING THE ACCURACY OF DNA BASE IDENTIFICATIONS".

En algunas realizaciones, los métodos pueden comprender las etapas de: (a) calcular un conjunto de valores de predictor para la identificación de bases; (b) usar los valores de predictor para buscar una puntuación de calidad en una tabla de calidad. En ciertas realizaciones, se realiza puntuación de calidad calculando un conjunto de predictores para cada identificación de bases, y usando esos valores de predictor para buscar la puntuación de calidad en una tabla de calidad. En algunas realizaciones, la tabla de calidad se genera usando una modificación del algoritmo de Phred en un conjunto de datos de calibración representativos de variabilidad de ejecución y secuencia. Los valores de predictor para cada identificación de bases puede ser cualquier aspecto adecuado que puede indicar o predecir la calidad de la identificación de bases en una ejecución de secuenciación dada. Por ejemplo, algunos predictores adecuados se exponen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537 presentada el 13 de enero de 2011 y titulado, "DATA PROCESSING SYSTEM AND METHODS". Como se describe en mayor detalle más adelante en este documento, valores de predictor adecuados pueden incluir, por ejemplo: solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; solapamiento de señal con fondo (SOWB); y ajuste de pureza G desplazada. Puede usarse cualquier combinación adecuada de los anteriores valores de predictor en los métodos presentados en este documento.

En ciertas realizaciones, los predictores de calidad usados en el algoritmo de Phred incluyen solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; y solapamiento de señal con fondo (SOWB).

Como se usa en este documento, "solapamiento en línea" se refiere a una medición de la separación entre las intensidades identificadas de primer plano y las intensidades de fondo. Por ejemplo, en algunas realizaciones, esta puntuación es una estadística que mide la señal a ruido de la lectura hasta la identificación de bases puntuada, y se pondera para tener más en cuenta las últimas identificaciones de bases, aunque incluso las primeras identificaciones de bases en la lectura tienen una influencia.

Como se usa en este documento, "pureza" se refiere a una medición que captura la fiabilidad de la probabilidad de que una identificación de base se base únicamente en el ciclo actual, y mide cómo de significativa es la base identificada cuando se compara con las otras tres bases.

Como se usa en este documento, "ajuste de fase" se refiere a una medición del ruido transportado desde los ciclos anterior y siguiente, que es esencialmente la suma de pesos de ajuste y preajuste de fase.

Como se usa en este documento, "Start5" se refiere a una métrica binaria que captura la fragmentación de preparación de muestras en el comienzo de una lectura. Por ejemplo, en una realización ilustrativa, este predictor puede recibir una puntuación binaria de "1" durante los primeros 5 ciclos, y "0" para cada ciclo posteriormente.

Como se usa en este documento, "puntuación de hexámeros" se refiere a una medición que examina hexámeros y devuelve un factor de enriquecimiento que refleja cuánto se enriquece el hexámero cerca de errores específicos de secuencia. Por ejemplo, en algunas realizaciones, esta puntuación asocia una medida de dificultad de secuenciación a cada secuencia de seis bases y se aplica comenzando en el ciclo 6 de la ejecución. Por lo tanto, los valores aplicados antes del ciclo 6 son el valor medio del predictor cuando todos los hexámeros se promedian juntos.

Como se usa en este documento, "acumulación de motivos" se refiere a una medición que mantiene una suma acumulativa del predictor de puntuación de hexámeros, que tiene en cuenta cómo de difícil ha sido el contexto de secuencia en los ciclos anteriores de la lectura. Por ejemplo, en algunas realizaciones, esta puntuación es la suma acumulativa de la puntuación de hexámeros y se concibe para medir la dificultad general de la lectura de secuenciación hasta la identificación de bases puntuada.

Como se usa en este documento, "resistencia" se refiere a una medición que rastrea lo cerca que la lectura está de finalización. Por ejemplo, en algunas realizaciones, esta puntuación es la recíproca del número de ciclo.

Como se usa en este documento, "homopolímero aproximado" se refiere a un cálculo del número de identificaciones de bases idénticas consecutivas que preceden a una identificación de bases. En ciertas realizaciones, el cálculo puede permitir una excepción, para identificar contextos de secuencia problemáticos tales como ejecuciones de homopolímeros y motivos problemáticos tales como "GGCGG".

Como se usa en este documento, "decaimiento de intensidad" se refiere a la identificación de identificaciones de bases que sufren pérdida de señal a medida que la secuenciación progresa. Por ejemplo, esto puede hacerse comparando la intensidad más brillante en el ciclo actual con la intensidad más brillante en el ciclo 1.

Como se usa en este documento, "penúltima depuración" se refiere a una medición de calidad de lectura temprana en las primeras 25 bases basándose en el segundo peor valor de depuración. Por ejemplo, en algunas realizaciones, esta puntuación se refiere a la calidad de lectura, que se correlaciona con el nivel de calidad general en los primeros 25 ciclos. Este predictor es muy similar a los criterios usados para marcar una lectura como filtrada o no filtrada, y tiene el efecto de hacer las puntuaciones de calidad agnósticas en cuanto a si se analizan todos los datos de una ejecución o únicamente los datos que pasan el filtro. La depuración puede determinarse como el mayor valor de intensidad dividido por la suma del mayor valor de intensidad y el segundo mayor valor de intensidad, en la que los valores de intensidad se obtienen a partir de cuatro canales de color. Por ejemplo, en algunas realizaciones, métodos de evaluación de calidad pueden incluir adicionalmente identificar lecturas en las que la segunda peor depuración en el primer subconjunto de identificaciones de bases está por debajo de un umbral, y marcar esas lecturas como datos de mala calidad. El primer subconjunto de identificaciones de bases puede ser cualquier número adecuado de identificaciones de bases que proporciona un suficiente. Por ejemplo, el subconjunto puede ser el primer 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 o mayor que las primeras 25 identificaciones de bases. Esto puede denominarse filtrado de lectura, de tal forma que en ciertas realizaciones, grupos que cumplen este corte se denominan como que han "superado el filtro".

Como se usa en este documento, "solapamiento de señal con fondo" (SOWB) se refiere a una medición de la separación de la señal del ruido en los ciclos anteriores y posteriores. En una realización preferida, la medición utiliza los 5 ciclos que preceden y siguen inmediatamente al ciclo actual.

Como se usa en este documento, "ajuste de pureza desplazada G" se refiere a una medición de la separación de la señal del ruido para únicamente la identificación de bases actual, mientras también tiene en cuenta los efectos atenuantes G. Debido a una interacción entre el colorante y la base de ADN incorporada en el ciclo anterior, las intensidades en ciertos canales de color pueden disminuirse (atenuarse) en ciclos después de esos ciclos en los que se incorporó un nucleótido G.

Después de calcular puntuaciones de calidad, pueden realizarse opcionalmente operaciones adicionales. Por lo tanto, en algunas realizaciones, el método para evaluar la calidad de una identificación de bases comprende además descontar puntuaciones de calidad no fiables al final de cada lectura. En realizaciones preferidas, la etapa de descontar puntuaciones de calidad no fiables comprende usar un algoritmo para identificar un umbral de fiabilidad. En una realización más preferida, identificaciones de bases fiables comprenden valores q por encima del umbral e identificaciones de bases no fiables comprenden valores q por debajo del umbral. Un algoritmo para determinar un umbral de fiabilidad puede comprender el algoritmo de Segmentos de Puntuación Máxima de Extremo Anclado (EAMSS), por ejemplo. Como se usa en este documento, un "algoritmo EAMSS" es un algoritmo que identifica puntos de transición en los que identificaciones de bases buenas y fiables (mayoritariamente con altos valores q) se convierten en identificaciones de bases no fiables (mayoritariamente con bajos valores q). La identificación de tales puntos de transición puede hacerse, por ejemplo, usando un Modelo Oculto de Markov que identifica desplazamientos en las distribuciones locales de puntuaciones de calidad. Por ejemplo, puede usarse un Modelo Oculto de Markov. Modelos Ocultos de Markov útiles se describen, por ejemplo, en Lawrence R. Rabiner (febrero 1989). "A tutorial on Hidden Markov Models and selected applications in speech recognition". Actas del IEEE 77 (2): 257-286. doi: 10.1109/5.18626. Sin embargo, será evidente para un experto en la materia que puede emplearse cualquier método adecuado de descuento de puntuaciones de calidad no fiables. En una realización preferida, identificaciones de bases no fiables pueden incluir identificaciones de bases con un sesgo fuerte hacia identificaciones de bases G.

Métricas en tiempo real

Los métodos y sistemas proporcionados en este documento también pueden utilizar métricas en tiempo real para visualizar la calidad de ejecución a un usuario. Las métricas pueden visualizarse como gráficos, diagramas, tablas, instantáneas o cualquier otro método de visualización adecuado que proporciona una representación significativa o útil de algún aspecto de calidad de ejecución a un usuario. Por ejemplo, las métricas en tiempo real visualizadas a un usuario pueden incluir una visualización de valores de intensidad en los ciclos de una ejecución, la calidad del enfoque de equipo óptico y densidad de grupos en cada carril. Visualizaciones de métricas adicionales pueden incluir puntuación Q , mostrada como una distribución basándose en la puntuación Q , o como un mapa de calor sobre una base de por ciclo, por ejemplo. En algunas realizaciones, las métricas en tiempo real pueden incluir una tabla resumen de diversos parámetros, clasificados por, por ejemplo, carril, placa o número de ciclo. Datos de imagen de una placa entera o subregión de una placa pueden visualizarse para una confirmación visual de calidad de imagen. Tales datos de imagen pueden incluir primer plano, imágenes en miniatura de alguna o todas las partes de una imagen.

Adicionalmente, algunas visualizaciones de métricas pueden incluir la tasa de error sobre una base de por ciclo. La tasa de error puede calcularse usando un ácido nucleico de control.

Métodos de secuenciación

Los métodos descritos en este documento pueden usarse en conjunto con diversas técnicas de secuenciación de ácido nucleico. Particularmente técnicas aplicables son en las que los ácidos nucleicos se unen a ubicaciones fijas en una matriz de tal forma que sus posiciones relativas no cambian y en las que la matriz se representa repetidamente. Son particularmente aplicables realizaciones en las que se obtienen imágenes en diferentes canales de color, por ejemplo, que coinciden con diferentes etiquetas usadas para distinguir un tipo de base de nucleótido de otro. En algunas realizaciones, el proceso para determinar la secuencia de nucleótido de un ácido nucleico objetivo puede ser un proceso automatizado. Realizaciones preferidas incluyen técnicas de secuenciación por síntesis ("SBS").

Las técnicas de SBS generalmente implican la extensión enzimática de una cadena de ácido nucleico naciente a través de la adición iterativa de nucleótidos contra una cadena molde. En métodos tradicionales de SBS, puede proporcionarse un monómero de un solo nucleótido a un nucleótido objetivo en la presencia de una polimerasa en cada suministro. Sin embargo, en los métodos descritos en este documento, puede proporcionarse más de un tipo de monómero de nucleótido a un ácido nucleico objetivo en la presencia de una polimerasa en un suministro.

SBS puede utilizar monómeros de nucleótido que tienen una fracción de terminación o los que carecen de cualquier fracción de terminación. Métodos que utilizan monómeros de nucleótido que carecen de terminadores incluyen, por ejemplo, pirosecuenciación y secuenciación usando nucleótidos etiquetados con γ -fosfato, como se expone en detalle adicional a continuación. En métodos que usan monómeros de nucleótido que carecen de terminadores, el número de nucleótidos añadidos en cada ciclo es generalmente variable y depende de la secuencia molde y el modo de suministro de nucleótido. Para técnicas de SBS que utilizan monómeros de nucleótido que tienen una fracción de terminación, el terminador puede ser irreversible de forma efectiva en las condiciones de secuenciación usadas como en el caso para la secuenciación de Sanger tradicional que utiliza didesoxinucleótidos, o el terminador puede ser reversible como es el caso para los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

Técnicas de SBS pueden utilizar monómeros de nucleótido que tienen una fracción de etiqueta o los que carecen de una fracción de etiqueta. Por consiguiente, pueden detectarse eventos de incorporación basándose en una característica de la etiqueta, tal como fluorescencia de la etiqueta; una característica del monómero de nucleótido tal como peso molecular o carga; un subproducto de la incorporación del nucleótido, tal como liberación de pirofosfato; o similares. En las realizaciones, en las que dos o más nucleótidos diferentes están presentes en un reactivo de secuenciación, los nucleótidos diferentes pueden ser distinguibles entre sí, o como alternativa, las dos o más etiquetas diferentes pueden ser las indistinguibles en las técnicas de detección que se usan. Por ejemplo, los diferentes nucleótidos presentes en un reactivo de secuenciación pueden tener diferentes etiquetas y pueden distinguirse usando ópticas apropiadas como se ilustra por los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

Realizaciones preferidas incluyen técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPi) a medida que se incorporan nucleótidos particulares en la cadena naciente (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. y Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release". *Analytical Biochemistry* 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing". *Genome Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. y Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate". *Science* 281(5375), 363; Patente de Estados Unidos N.º 6.210.891; Patente de Estados Unidos N.º 6.258.568 y Patente de Estados Unidos N.º 6.274.320). En la pirosecuenciación, el PPi liberado puede detectarse convirtiéndose inmediatamente en trifosfato de adenosina (ATP) mediante ATP sulfurilasa, y el nivel de ATP generado se detecta a través de fotones producidos por luciferasa. Los ácidos nucleicos que hay que secuenciar pueden unirse a características en una matriz y la matriz puede representarse para capturar las señales quimioluminiscentes que se producen debido a la incorporación de un nucleótido a las características de la matriz. Puede obtenerse una imagen después de que la matriz se trata con un tipo de nucleótido particular (por ejemplo, A, T, C o G). Las imágenes obtenidas después de la adición de cada tipo de nucleótido diferirán con respecto a qué características se detectan en la matriz. Estas diferencias en la imagen reflejan el diferente contenido de secuencia de las características en la matriz. Sin embargo, las ubicaciones relativas de cada característica permanecerán sin cambios en las imágenes. Las imágenes pueden almacenarse, procesarse y analizarse usando los métodos expuestos en este documento. Por ejemplo, imágenes obtenidas después del tratamiento de la matriz con cada diferente tipo de nucleótido pueden tratarse de la misma manera que se ilustra en este documento para imágenes obtenidas de diferentes canales de detección para métodos de secuenciación basada en terminadores reversibles.

En otro tipo ilustrativo de SBS, se consigue secuenciación de ciclo mediante adición escalonada de nucleótidos de terminación reversible que contienen, por ejemplo, una etiqueta de colorante escindible o fotoblanqueable como se describe, por ejemplo, en el documento WO 04/018497 y la Patente de Estados Unidos N.º 7.057.026. Este enfoque se está comercializando por Solexa (actualmente Illumina Inc.), y se describe también en el documento WO 91/06678 y el documento WO 07/123.744. La disponibilidad de terminadores etiquetados fluorescentemente en la que tanto la terminación puede reservarse como la etiqueta fluorescente escindible facilita la secuenciación de terminación reversible cíclica (CRT). También pueden diseñarse polimerasas para incorporar de forma eficiente y extender a partir de estos nucleótidos modificados.

Preferentemente en realizaciones de secuenciación basada en terminadores reversibles, las etiquetas sustancialmente no inhiben la extensión en condiciones de reacciones de SBS. Sin embargo, las etiquetas de detección pueden ser extraíbles, por ejemplo, por escisión o degradación. Pueden capturarse imágenes después de la incorporación de etiquetas en características de ácido nucleico en matriz. En realizaciones particulares, cada ciclo implica el suministro simultáneo de cuatro tipos de nucleótidos diferentes a la matriz y cada tipo de nucleótido tiene una etiqueta espectralmente distinta. A continuación pueden obtenerse cuatro imágenes, usando cada una un canal de detección que es selectivo para una de las cuatro etiquetas diferentes. Como alternativa, pueden añadirse secuencialmente diferentes tipos de nucleótidos y puede obtenerse una imagen de la matriz entre cada etapa de adición. En tales realizaciones cada imagen mostrará características de ácido nucleico que han incorporado nucleótidos de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica. Sin embargo, la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de tales métodos de SBS de terminadores reversibles pueden almacenarse, procesarse y analizarse como se expone en este documento. Después de la etapa de captura de imagen, pueden eliminarse etiquetas y pueden eliminarse fracciones de terminación reversibles para ciclos posteriores de adición y detección de nucleótidos. La eliminación de las etiquetas después de que se hayan detectado en un ciclo particular y antes de un ciclo posterior puede proporcionar la ventaja de reducir señal de fondo y diafonía entre ciclos. A continuación se exponen ejemplos de etiquetas útiles y métodos de eliminación.

En realizaciones particulares algunos o todos los monómeros de nucleótido pueden incluir terminadores reversibles. En tales realizaciones, los terminadores reversibles/flúores escindibles pueden incluir flúor enlazado a la fracción de ribosa a través de un enlace 3' éster (Metzker, *Genome Res.* 15:1767-1776 (2005)). Otros enfoques han separado la química de terminación de la escisión de la etiqueta de fluorescencia (Ruparel et al., *Proc Natl Acad Sci USA* 102: 5932-7 (2005)). Ruparel et al describieron el desarrollo de terminadores reversibles que usaban un pequeño grupo 3' alilo para bloquear la extensión, pero podría desbloquearse fácilmente mediante un tratamiento corto con un catalizador de paladio. El fluoróforo se unió a la base a través de un enlazante fotoescindible que podría escindirse fácilmente por una exposición de 30 segundos a luz UV de longitud de onda larga. Por lo tanto, puede usarse o bien reducción de disulfuro o bien fotoescisión como un enlazante escindible. Otro enfoque a terminación reversible es el uso de terminación natural que garantiza después de la colocación de un colorante voluminoso en un dNTP. La presencia de un colorante voluminoso cargado en el dNTP puede actuar como un terminador efectivo a través de impedimento estérico y/o electroestático. La presencia de un evento de incorporación evita incorporaciones adicionales a no ser que se elimine el colorante. La escisión del colorante elimina el flúor e invierte eficazmente la terminación. Ejemplos de nucleótidos modificados también se describen en la Patente de Estados Unidos N.º 7.427.673 y la Patente de Estados Unidos N.º 7.057.026.

Sistemas y métodos de SBS ilustrativos adicionales que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2007/0166705, Publicación de solicitud de Patente de Estados Unidos N.º 2006/0188901, Patente de Estados Unidos N.º 7.057.026, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0240439, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0281109, Publicación PCT N.º WO 05/065814, Publicación de Solicitud de Patente de Estados Unidos N.º 2005/0100900, Publicación PCT N.º WO 06/064199, Publicación PCT N.º WO 07/010.251, Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0270305 y Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0260372.

Algunas realizaciones pueden utilizar detección de cuatro diferentes nucleótidos usando menos de cuatro etiquetas diferentes. Por ejemplo, puede realizarse SBS utilizando métodos y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, pueden detectarse un par de tipos de nucleótidos en la misma longitud de onda, pero distinguirse basándose en una diferencia en intensidad para un miembro del par en comparación con el otro, o basándose en un cambio a un miembro del par (por ejemplo, a través de modificación química, modificación fotoquímica o modificación física) que provoca que la señal aparente aparezca o desaparezca en comparación con la señal detectada para el otro miembro del par. Como un segundo ejemplo, pueden detectarse tres de cuatro diferentes tipos de nucleótidos en condiciones particulares mientras un cuarto tipo de nucleótido carece de una etiqueta que es detectable en esas condiciones, o se detecta mínimamente en esas condiciones (por ejemplo, detección mínima debido a fluorescencia de fondo, etc.). La incorporación de los primeros tres tipos de nucleótidos en un ácido nucleico puede determinarse basándose en la presencia de sus respectivas señales y la incorporación del cuarto tipo de nucleótido en el ácido nucleico puede determinarse basándose en la ausencia o detección mínima de cualquier señal. Como un tercer ejemplo, un tipo de nucleótido puede incluir etiqueta o etiquetas que se detectan en dos canales diferentes, mientras que otros tipos de nucleótidos se detectan en no más de uno de los canales. Las tres configuraciones ilustrativas anteriormente mencionadas no se consideran mutuamente exclusivas y pueden usarse en diversas combinaciones. Una realización ilustrativa que combina los tres ejemplos, es un método de SBS basado en fluorescencia que usa un primer tipo de nucleótido que se detecta en un primer canal (por ejemplo, dATP que tiene una etiqueta que se detecta en el primer canal cuando se excita por una primera longitud de onda de excitación), un segundo tipo de nucleótido que se detecta en un segundo canal (por ejemplo, dCTP que tiene una etiqueta que se detecta en el segundo canal cuando se excita por una segunda longitud de onda de excitación), un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal (por ejemplo, dTTP que tiene al menos una etiqueta que se detecta en ambos canales cuando se excita por la primera y/o segunda longitud

de onda de excitación) y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o se detecta mínimamente, en cada canal (por ejemplo, dGTP que no tiene etiqueta).

Además, como se describe en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232, pueden obtenerse datos de secuenciación usando un único canal. En tales así llamados enfoques de secuenciación de un colorante, el primer tipo de nucleótido se etiqueta pero la etiqueta se elimina después de que se genera la primera imagen, y el segundo tipo de nucleótido se etiqueta únicamente después de que se genera una primera imagen. El tercer tipo de nucleótido retiene su etiqueta tanto en la primera como segunda imágenes, y el cuarto tipo de nucleótido permanece sin etiquetar en ambas imágenes.

Algunas realizaciones pueden utilizar secuenciación mediante técnicas de ligado. Tales técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de tales oligonucleótidos. Los oligonucleótidos habitualmente tienen diferentes etiquetas que se correlacionan con la identidad de un nucleótido particular en una secuencia a la que los oligonucleótidos hibridan. Como con otros métodos de SBS, las imágenes pueden obtenerse después del tratamiento de una matriz de características de ácido nucleico con los reactivos de secuenciación etiquetados. Cada imagen mostrará características de ácido nucleico que han incorporado etiquetas de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica, pero la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de métodos de secuenciación basados en ligado pueden almacenarse, procesarse y analizarse como se expone en este documento. Sistemas y métodos de SBS ilustrativos que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Patente de Estados Unidos N.º 6.969.488, Patente de Estados Unidos N.º 6.172.218, y Patente de Estados Unidos N.º 6.306.597.

Algunas realizaciones pueden utilizar secuenciación por nanoporos (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing". Trends Biotechnol. 18, 147-151 (2000); Deamer, D. Y D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, y J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). En tales realizaciones, el ácido nucleico objetivo pasa a través de un nanoporo. El nanoporo puede ser un poro sintético o proteína de membrana biológica, tal como α -hemolisina. Como el ácido nucleico objetivo pasa a través del nanoporo, cada par de bases puede identificarse midiendo fluctuaciones en la conductancia eléctrica del poro. (Patente de Estados Unidos N.º 7.001.792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores". Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis". Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution". J. Am. Chem. Soc. 130, 818-820 (2008)). Pueden almacenarse, procesarse y analizarse datos obtenidos de secuenciación por nanoporos como se expone en este documento. En particular, los datos pueden tratarse como una imagen de acuerdo con el tratamiento ilustrativo de imágenes ópticas y otras imágenes que se expone en este documento.

Algunas realizaciones pueden utilizar métodos que implican la supervisión en tiempo real de actividad de ADN polimerasa. Pueden detectarse incorporaciones de nucleótidos a través de interacciones de transferencia de energía de resonancia de fluorescencia (FRET) entre una polimerasa que porta fluoróforo y nucleótidos etiquetados con γ -fosfato como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.329.492 y Patente de Estados Unidos N.º 7.211.414 o pueden detectarse incorporaciones de nucleótidos con guías de ondas en modo cero como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.315.019 y usando análogos de nucleótidos fluorescentes y polimerasas diseñadas como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.405.281 y Publicación de Solicitud de Patente de Estados Unidos N.º 2008/0108082. La iluminación puede restringirse a un volumen de escala de zeptolitro alrededor de una polimerasa anclada a superficie de tal forma que la incorporación de nucleótidos etiquetados fluorescentemente puede observarse con poco fondo (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations". Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time". Opt. Lett. 33, 1026-1028 (2008); Koriach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures". Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). Las imágenes obtenidas a partir de tales métodos pueden almacenarse, procesarse y analizarse como se expone en este documento.

Los métodos de SBS anteriores pueden efectuarse ventajosamente en formatos de multiplex de tal forma que se manipulan simultáneamente múltiples ácidos nucleicos objetivo diferentes. En realizaciones particulares, diferentes ácidos nucleicos objetivo pueden tratarse en un recipiente de reacción común o una superficie de un sustrato particular. Esto permite el suministro de reactivos de secuenciación, eliminación de reactivos sin reaccionar y detección de eventos de incorporación de una manera multiplex. En las realizaciones que usan ácidos nucleicos objetivo enlazados a superficie, los ácidos nucleicos objetivo pueden ser en un formato de matriz. En un formato de matriz, los ácidos nucleicos objetivo pueden enlazarse habitualmente a una superficie de una manera espacialmente distinguible. Los ácidos nucleicos objetivo pueden enlazarse por unión covalente directa, unión a una perla u otra partícula o enlace a una polimerasa u otra molécula que se une a la superficie. La matriz puede incluir una única copia de un ácido nucleico objetivo en cada sitio (también denominado como una característica) o múltiples copias que tiene la misma secuencia pueden estar presentes en cada sitio o característica. Múltiples copias pueden producirse mediante métodos de amplificación tales como, amplificación de puente o PCR en emulsión como se describe en detalle adicional a

continuación.

Los métodos expuestos en este documento pueden usar matrices que tienen características en cualquiera de diversas densidades incluyendo, por ejemplo, al menos aproximadamente 10 características/cm², 100 características/cm², 500 características/cm², 1.000 características/cm², 5.000 características/cm², 10.000 características/cm², 50.000 características/cm², 100.000 características/cm², 1.000.000 características/cm², 5.000.000 características/cm², o mayores.

Se apreciará que cualquiera de los procesos de secuenciación anteriormente descritos puede incorporarse en los métodos y/o sistemas descritos en este documento. Adicionalmente, se apreciará que pueden implementarse fácilmente otros procesos de secuenciación conocidos para su uso con los métodos y/o sistemas descritos en este documento. Se apreciará también que los métodos y sistemas descritos en este documento se diseñan para ser aplicables con cualquier tecnología de secuenciación de ácido nucleico. Adicionalmente, se apreciará que los métodos y sistemas descritos en este documento tiene incluso una aplicabilidad más amplia para cualquier campo en el que el rastreo y análisis de características en un espécimen con el paso del tiempo o desde diferentes perspectivas es importante. Por ejemplo, los métodos y sistemas descritos en este documento pueden aplicarse donde datos de imagen obtenidos por tecnologías de formación de imágenes de vigilancia, aéreas o por satélite y similares se adquieren en puntos de tiempo o perspectivas diferentes y analizan.

Sistemas

Un sistema capaz de efectuar un método expuesto en este documento, si se integra con capacidades de detección o no, puede incluir un controlador de sistema que es capaz de ejecutar un conjunto de instrucciones para realizar una o más etapas de un método, técnica o proceso expuesto en este documento. Por ejemplo, las instrucciones pueden dirigir la realización de etapas para crear un conjunto de amplicones *in situ*. Opcionalmente, las instrucciones pueden dirigir adicionalmente la realización de etapas para detectar ácidos nucleicos usando métodos expuestos anteriormente en este documento. Un controlador de sistema útil puede incluir cualquier sistema basado en procesador o basado en microprocesador, incluyendo sistemas que usan microcontroladores, ordenadores de conjunto reducido de instrucciones (RISC), circuitos integrados específicos de la aplicación (ASIC), campos de matriz de puertas programables (FPGA), circuitos lógicos y cualquier otro circuito o procesador capaz de ejecutar funciones descritas en este documento. Un conjunto de instrucciones para un controlador de sistema puede ser en forma de un programa de software. Como se usa en este documento, los términos "software" y "firmware" son intercambiables, e incluyen cualquier programa informático almacenado en memoria para su ejecución por un ordenador, incluyendo memoria RAM, memoria ROM, memoria EPROM, memoria EEPROM y memoria RAM no volátil (NVRAM). El software puede ser en diversas formas tales como software de sistema o software de aplicación. Además, el software puede ser en forma de una colección de programas separados, o un módulo de programa dentro de un programa mayor o una porción de un módulo de programa. El software también puede incluir programación modular en forma de programación orientada a objeto.

El término comprendiendo/que comprende se concibe en este documento para ser abierto, incluyendo no únicamente los elementos citados, sino incluyendo adicionalmente cualquier elemento adicional.

Un número de las realizaciones se han descrito. Sin embargo, se entenderá que pueden hacerse diversas modificaciones.

La siguiente descripción es con respecto a las Figuras 1A, 1B y 2. Realizaciones descritas en lo sucesivo también se describen en la Solicitud Provisional de Estados Unidos N.º 61/915.455, presentada el jueves, 12 de diciembre de 2013.

El análisis de datos de imagen presenta un número de retos, especialmente con respecto a la comparación de imágenes de un artículo o estructura que se capturan desde diferentes puntos de referencia. La mayoría de metodología de análisis de imágenes emplea, al menos en parte, etapas para alinear múltiples imágenes separadas entre sí basándose en características o elementos presentes en ambas imágenes. Diversas realizaciones de las composiciones y métodos divulgados en este documento mejoran métodos anteriores para el análisis de imágenes. Algunos métodos anteriores para análisis de imágenes se exponen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537 presentada en 13 de enero de 2011 y titulada, "DATA PROCESSING SYSTEM AND METHODS".

En la actualidad, se han desarrollado herramientas que adquieren y analizan datos de imagen generados en puntos de tiempo o perspectivas diferentes. Algunos ejemplos incluyen herramientas para análisis de imágenes de satélite y herramientas de biología molecular para secuenciar y caracterizar la identidad molecular de un espécimen. En cualquier tal sistema, adquirir y almacenar grandes números de imágenes de alta calidad habitualmente requiere cantidades masivas de capacidad de almacenamiento. Adicionalmente, una vez adquiridos y almacenados, el análisis de datos de imagen puede volverse intensivo en recursos y puede interferir con capacidad de procesamiento de otras funciones, tales como adquisición constante y almacenamiento de datos de imagen adicionales. Como tal, serían beneficiosos métodos y sistemas que mejoren la velocidad y precisión de análisis de la adquisición y análisis de datos de imagen.

En el campo de la biología molecular, uno de los procesos para secuenciación de ácido nucleico en uso es secuenciación por síntesis. La técnica puede aplicarse a proyectos de secuenciación masivamente paralelos. Por ejemplo, usando una plataforma automatizada, es posible efectuar cientos de miles de reacciones de secuenciaciones simultáneamente. Por lo tanto, la divulgación se refiere a instrumentos y métodos para adquirir, almacenar y analizar datos de imagen generados durante secuenciación de ácido nucleico.

Enormes ganancias en la cantidad de datos que pueden adquirirse y almacenarse hacen incluso más beneficiosos a los métodos de análisis de imagen simplificados. Por ejemplo, los métodos de análisis de imagen descritos en este documento permiten que tanto diseñadores como usuarios finales hagan un uso eficiente de hardware informático existente. Por consiguiente, en este documento se presentan métodos y sistemas que reducen la carga de cálculo del procesamiento de datos ante el rápido crecimiento de la emisión de datos. Por ejemplo, en el campo de secuenciación de ADN, los rendimientos han escalado 15 veces durante el transcurso del último año, y ahora puede alcanzar cientos de gigabases en una única ejecución de un dispositivo de secuenciación de ADN. Si los requisitos de infraestructura de cálculo crecen exponencialmente, los experimentos de genoma a gran escala permanecerían fuera del alcance de la mayoría de investigadores. Por lo tanto, la generación de más datos de secuencia sin procesar aumentará la necesidad de análisis secundario y almacenamiento de datos, haciendo extremadamente valiosa la optimización de transporte y almacenamiento de datos. Algunas realizaciones de los métodos y sistemas presentados en este documento pueden reducir el tiempo, hardware, interconexión de redes y requisitos de infraestructura de laboratorio necesarios para producir datos de secuencia usables.

Como se usa en este documento, una "característica" es un área de interés dentro de un espécimen o campo de visión. Cuando se usa en conexión con dispositivos de micromatriz u otros dispositivos de analítica molecular, una característica se refiere al área ocupada por moléculas similares o idénticas. Por ejemplo, una característica puede ser un oligonucleótido amplificado o cualquier otro grupo de un polinucleótido o polipéptido con una misma o similar secuencia. En otras realizaciones, una característica puede ser cualquier elemento o grupo de elementos que ocupa un área física en un espécimen. Por ejemplo, una característica podría ser una parcela de tierra, una masa de agua o similar. Cuando se representa una característica, cada característica tendrá alguna área. Por lo tanto, en muchas realizaciones, una característica no es meramente un píxel.

Las distancias entre características pueden describirse de cualquier número de formas. En algunas realizaciones, las distancias entre características pueden describirse a partir del centro de una característica hasta el centro de otra característica. En otras realizaciones, las distancias pueden describirse desde el borde de una característica hasta el borde de otra característica, o entre los puntos identificables más externos de cada característica. El borde de una característica puede describirse como el límite físico teórico o real en un chip, o algún punto dentro del límite de la característica. En otras realizaciones, las distancias pueden describirse en relación con un punto fijo en el espécimen o en la imagen del espécimen.

Pueden secuenciarse múltiples copias de ácidos nucleicos en una característica, por ejemplo, proporcionando una base de nucleótido etiquetada a la matriz de moléculas, extendiendo de este modo un cebador hibridado a un ácido nucleico dentro de una característica para producir una señal que corresponde a una característica que comprende el ácido nucleico. En realizaciones preferidas, los ácidos nucleicos dentro de una característica son idénticos o sustancialmente idénticos entre sí.

En algunos de los métodos de análisis de imagen descritos en este documento, cada imagen en el conjunto de imágenes incluye señales de colores, en el que un color diferente corresponde a una base de nucleótido diferente. En algunos aspectos, cada imagen del conjunto de imágenes comprende señales que tienen un único color seleccionado de al menos cuatro colores diferentes. En ciertos aspectos, cada imagen en el conjunto de imágenes comprende señales que tienen un único color seleccionado de cuatro colores diferentes.

Con respecto a ciertos métodos de cuatro canales descritos en este documento, pueden secuenciarse ácidos nucleicos proporcionando, cuatro diferentes bases de nucleótidos etiquetadas a la matriz de moléculas para producir cuatro imágenes diferentes, comprendiendo cada imagen señales que tienen un único color, en el que el color de señal es diferente para cada una de las cuatro imágenes diferentes, produciendo de este modo un ciclo de cuatro imágenes de color que corresponde a los cuatro posibles nucleótidos presentes en una posición particular en el ácido nucleico. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de color.

Con respecto a ciertos métodos de dos canales descritos en este documento, pueden secuenciarse ácidos nucleicos utilizando métodos y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, puede secuenciarse un ácido nucleico proporcionando un primer tipo de nucleótido que se detecta en un primer canal, un segundo tipo de nucleótido que se detecta en un segundo canal, un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o está mínimamente, en ningún canal. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de

moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de color.

Estimación de ajuste de fase

5 Una estimación de ajuste de fase es una herramienta analítica para reducir el ruido durante múltiples ciclos en una ejecución de secuenciación. Por ejemplo, en cualquier ciclo dado de una ejecución de secuencia, una o más moléculas pueden "ajustarse en fase" en cada ciclo. Como se usa en este documento, "ajustar en fase", "ajuste de fase" y términos similares se refieren a la situación en la que una molécula en una característica cae al menos una base por debajo de otras moléculas en la misma característica como resultado de que la característica se secuencie en un ciclo particular. Como se usa en este documento, "preajustar en fase", "preajuste de fase" y términos similares se refieren a la situación en la que una molécula en una característica salta al menos una base por delante de otras moléculas en la misma característica como resultado de que la característica se secuencie un ciclo particular. Los efectos de ajuste de fase y preajuste de fase se vuelven más pronunciados con mayores tasas de ajuste/preajuste de fase y lecturas más largas. Por lo tanto, para mantener una identificación de bases precisa durante un número de ciclos extendidos, es importante corregir para este fenómeno. Los métodos y sistemas presentados en este documento proporcionan una solución de cálculo que sorpresivamente produce una identificación de bases mejorada durante ciclos de secuenciación extendidos en comparación con métodos de corrección de ajuste de fase tradicionales.

20 Los métodos y sistemas proporcionados en este documento pueden asumir que una fracción fija de moléculas en cada característica se ajusta en fase en cada ciclo, en el sentido de que esas moléculas caen una base por detrás en secuenciación. Por lo tanto, en una realización preferida, se realiza una estimación de ajuste de fase para ajustar las intensidades observadas de una forma que reduce el ruido creado por moléculas ajustadas en fase.

25 Puede realizarse corrección de ajuste de fase tradicional mediante métodos como se describen en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537. Como se describe en la misma, un enfoque tradicional a corrección de ajuste de fase implica crear una matriz de ajuste de fase para modelar los efectos de ajuste de fase en cualquier ciclo dado. Esto puede hacerse, por ejemplo, creando una matriz NxN en la que N es el número total de ciclos. A continuación, para corregir con ajuste de fase intensidades para un ciclo dado, se toma la inversa de la matriz de ajuste de fase y se extrae la fila de matriz que corresponde al ciclo. Como resultado, el vector de intensidades reales para ciclos 1 a N es producto de la inversa de matriz de ajuste de fase e intensidades observadas para ciclos 1 a N. Como un ejemplo de un enfoque de este tipo, se realiza una estimación de ajuste de fase calculando tasas ajuste de fase y preajuste de fase a partir de los primeros 12 ciclos de datos de intensidad. Correcciones derivadas de estas tasas se aplican a continuación a todos los ciclos para mejorar tasas de error de identificación de bases. Debido a que las tasas de ajuste de fase se estiman durante la parte temprana de una ejecución de secuencia, una estimación imprecisa de tasa de ajuste de fase hecha durante ciclos tempranos (por ejemplo, durante los ciclos 1-12) puede afectar potencialmente a los datos obtenidos durante ciclos posteriores.

40 Por ejemplo, en métodos corrección de ajuste de fase tradicionales, si la estimación de tasa de ajuste de fase está apagada, la precisión de identificación de bases se ve afectada para la totalidad de una ejecución y no se ajusta. Este efecto se mejora cuando se secuencian muestras de diversidad baja tales como amplicones individuales. Por lo tanto, si tasas de ajuste de fase estimadas durante ciclos tempranos se basan en una baja diversidad de bases, las tasas pueden no reflejar con precisión las tasas de ajuste de fase durante ciclos posteriores de una ejecución de secuencia. Enfoques de corrección de ajuste de fase tradicionales no son efectivos en ajustar a tasas de ajuste de fase cambiantes en ciclos posteriores. Adicionalmente, los enfoques de corrección de ajuste de fase tradicionales no están diseñados para estimar la tasa de ajuste de fase en datos de 2 canales.

Corrección de ajuste de fase empírica

50 En este documento se presentan métodos mejorados de realización de corrección de ajuste de fase. Los métodos descritos en este documento proporcionan ventajas sorprendentes en comparación con los enfoques de corrección de ajuste de fase tradicionales descritos anteriormente. Por ejemplo, los métodos presentados en este documento incluyen determinar correcciones de ajuste de fase como un análisis en curso a lo largo de una ejecución de secuencia. Como resultado de este enfoque, estimación de tasa de ajuste de fase imprecisa hecha durante ciclos tempranos (por ejemplo, durante los ciclos 1-12) no afectará de forma adversa a ciclos posteriores.

55 Se presenta en este documento un método de realización de corrección de ajuste de fase que comprende análisis empírico. Los métodos presentados en este documento son una alternativa a, o pueden complementar los análisis de corrección de ajuste de fase tradicionales como se ha descrito anteriormente. Los métodos presentados en este documento son sorprendentemente efectivos cuando se aplican a, por ejemplo, datos de 1 canal y 2 canales.

60 En algunas realizaciones, los métodos comprenden una corrección de ajuste de fase empírica. Realizaciones particulares emplean la etapa de aplicar una corrección de ajuste de fase de primer orden. Por ejemplo, en algunas realizaciones, el método comprende una corrección de ajuste de fase de primer orden para un ciclo dado según se define mediante la siguiente:

$$I(\text{ciclo}) = I(\text{ciclo}) \times I(\text{ciclo}-1) \times I(\text{ciclo}+1)$$

en la que I representa intensidad y X e Y representan los pesos de ajuste y preajuste de fase calculados para este ciclo. Se entenderá que, utilizando este enfoque, si se eligen los valores correctos de X e Y , a continuación se maximiza la depuración media (calidad) de valores de intensidad. Por ejemplo, es posible optimizar numéricamente a través de una búsqueda de patrones en X e Y para maximizar la depuración media. Una vez que se identifican los valores X e Y con depuración media máxima, a continuación puede aplicarse la anterior corrección y, a continuación, puede producirse identificación de bases directamente posterior.

En algunas realizaciones, se calcula una corrección de ajuste de fase separada más de una vez durante una ejecución de secuencia. Por ejemplo, en algunas realizaciones, se calcula una corrección de ajuste de fase separada 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, o más de 100 veces durante una ejecución de secuencia. En algunas realizaciones, una corrección de ajuste de fase se calcula en casi cada ciclo durante una ejecución de secuencia. En algunas realizaciones, una corrección de ajuste de fase se calcula en cada ciclo durante una ejecución de secuencia.

En algunas realizaciones, se calcula una corrección de ajuste de fase separada para diferentes ubicaciones de una superficie representada en el mismo ciclo. Por ejemplo, en algunas realizaciones, se calcula una corrección de ajuste de fase separada para cada carril individual de una superficie representada, tal como un carril de célula de flujo individual. En algunas realizaciones se calcula una corrección de ajuste de fase separada para cada subconjunto de un carril, tal como una hilera de imagen dentro de un carril de célula de flujo. En algunas realizaciones, se calcula una corrección de ajuste de fase separada para cada imagen individual, tal como, por ejemplo, cada placa. En ciertas realizaciones, se calcula una corrección de ajuste de fase separada para cada placa en cada ciclo.

En realizaciones particulares, el enfoque descrito anteriormente para corrección de ajuste de fase empírica sirve para optimizar las correcciones de ajuste y preajuste de fase para cada ciclo y placa para maximizar la depuración media de los datos de intensidad. El resultado es que el RTA ya no depende de un cálculo de tasa preciso, ya que se aplica la mejor corrección en cada ciclo, pero en su lugar realiza correcciones ciclo a ciclo que se analizan en un ciclo posterior, por ejemplo, el ciclo 25. Este análisis proporciona una tasa calculada que puede guardarse en un archivo y/o visualizarse en una interfaz de usuario.

Como se expone en la Figura 1, los efectos de aplicación del enfoque anterior pueden resultar en una resolución dramática de identificación de bases. La Figura 1A muestra intensidades sin tratar para una placa particular y un ciclo particular en un sistema de dos canales en el que el nucleótido C se representa mediante señal en canal 1 únicamente, el nucleótido A se representa mediante señal en canal 2 únicamente, el nucleótido T se representa mediante señal en ambos canales 1 y 2, y el nucleótido G está "oscuro". La Figura 1B muestra intensidades con corrección de ajuste de fase de los mismos datos usando la corrección de ajuste de fase descrita anteriormente. Como se muestra en la Figura 1B, la aplicación del enfoque de corrección de ajuste de fase descrito anteriormente aumenta dramáticamente la resolución de intensidades asignadas a cada una de las cuatro bases. Para ayudar en la distinción de puntos de datos, los datos para los nucleótidos pueden indicarse en colores diferentes. Por ejemplo, los datos de nucleótido A pueden indicarse en verde, el nucleótido C puede indicarse en negro, el nucleótido T puede indicarse en rosa y el nucleótido G puede indicarse en azul.

En realizaciones particulares, debido a la física de ajuste de fase, a medida que las lecturas se vuelven más largas, términos de orden superior pueden volverse más y más importantes en corrección de ajuste de fase. Por lo tanto, en realizaciones particulares, para corregir esto, puede calcularse una corrección de ajuste de fase empírica de segundo orden. Por ejemplo, en algunas realizaciones, el método comprende una corrección de ajuste de fase de segundo orden como se define mediante la siguiente:

$$I(\text{ciclo}) = -a \cdot I(\text{ciclo}-2) + A \cdot I(\text{ciclo}-1) + I(\text{ciclo}) + B \cdot I(\text{ciclo}+1) - b \cdot I(\text{ciclo}+2)$$

en la que I representa intensidad y a , A , B y b representan los términos de primer y segundo orden a la corrección de ajuste de fase. En realizaciones particulares, el cálculo se optimiza en a , A , B , b .

En algunas realizaciones, pueden usarse términos de orden superior para corregir tasas de ajuste y/o preajuste de fase altas. En realizaciones particulares, cuanto más altas sean las tasas de ajuste y/o preajuste de fase, mayor será la diferencia que hace los términos de orden superior. En realizaciones particulares, cuanto más altas sean las tasas de ajuste y/o preajuste de fase y mayor sea la lectura, más importantes se volverán los términos de orden superior.

Los métodos proporcionados en este documento son superiores y proporcionan ventajas significativas sobre enfoques de corrección de ajuste de fase tradicionales. Por ejemplo, a diferencia de métodos tradicionales, no existe ningún requisito para estimar con precisión una tasa de ajuste de fase en los primeros 10 ciclos de una ejecución. Además, a diferencia de métodos tradicionales, no existe ningún requisito para agregar estimadas de ajuste de fase a través de placas para llegar a corrección de ajuste de fase que se generaliza a través de todas las placas. Además, a diferencia de métodos tradicionales en los que una corrección de ajuste de fase se deriva y aplica a todos los ciclos, en los métodos presentados en este documento, correcciones ciclo a ciclo son independientes. Específicamente, no se introduce error permanente en el algoritmo de corrección de ajuste de fase por unos pocos ciclos de datos malos.

Los métodos presentados en este documento no se ven afectados particularmente por ejecuciones de diversidad baja. Por ejemplo, en ejecuciones de secuenciación en las que únicamente se están determinando una o muy pocas secuencias, tal como en amplicón individual o en aplicaciones metagenómicas, la corrección de ajuste de fase no depende en su totalidad de la precisión de un cálculo hecha basándose en un conjunto limitado de ciclos tempranos, y en su lugar puede optimizar correcciones de ajuste de fase para cada placa y cada ciclo.

Aunque los métodos y sistemas presentados en este documento se ilustran esencialmente en el contexto de datos de secuenciación de dos canales, debería apreciarse que los mismos métodos y algoritmos pueden aplicarse directamente a datos de 4 canales con tasas de error sustancialmente reducidas en puntuaciones de alineaciones aumentadas. Un ejemplo de cálculos de corrección de ajuste de fase usando datos de 2 canales se presenta a continuación como Ejemplo 1. Un ejemplo de cálculos de corrección de ajuste de fase usando datos de 4 canales se presenta a continuación como Ejemplo 2.

Métodos de secuenciación

Los métodos descritos en este documento pueden usarse en conjunto con diversas técnicas de secuenciación de ácido nucleico. Particularmente técnicas aplicables son en las que los ácidos nucleicos se unen a ubicaciones fijas en una matriz de tal forma que sus posiciones relativas no cambian y en las que la matriz se representa repetidamente. Son particularmente aplicables realizaciones en las que se obtienen imágenes en diferentes canales de color, por ejemplo, que coinciden con diferentes etiquetas usadas para distinguir un tipo de base de nucleótido de otro. En algunas realizaciones, el proceso para determinar la secuencia de nucleótido de un ácido nucleico objetivo puede ser un proceso automatizado. Realizaciones preferidas incluyen técnicas de secuenciación por síntesis ("SBS").

Las técnicas de SBS generalmente implican la extensión enzimática de una cadena de ácido nucleico naciente a través de la adición iterativa de nucleótidos contra una cadena molde. En métodos tradicionales de SBS, puede proporcionarse un monómero de un solo nucleótido a un nucleótido objetivo en la presencia de una polimerasa en cada suministro. Sin embargo, en los métodos descritos en este documento, puede proporcionarse más de un tipo de monómero de nucleótido a un ácido nucleico objetivo en la presencia de una polimerasa en un suministro.

SBS puede utilizar monómeros de nucleótido que tienen una fracción de terminación o los que carecen de cualquier fracción de terminación. Métodos que utilizan monómeros de nucleótido que carecen de terminadores incluyen, por ejemplo, pirosecuenciación y secuenciación usando nucleótidos etiquetados con γ -fosfato, como se expone en detalle adicional a continuación. En métodos que usan monómeros de nucleótido que carecen de terminadores, el número de nucleótidos añadidos en cada ciclo es generalmente variable y depende de la secuencia molde y el modo de suministro de nucleótido. Para técnicas de SBS que utilizan monómeros de nucleótido que tienen una fracción de terminación, el terminador puede ser irreversible de forma efectiva en las condiciones de secuenciación usadas como en el caso para la secuenciación de Sanger tradicional que utiliza didesoxinucleótidos, o el terminador puede ser reversible como es el caso para los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

Técnicas de SBS pueden utilizar monómeros de nucleótido que tienen una fracción de etiqueta o los que carecen de una fracción de etiqueta. Por consiguiente, pueden detectarse eventos de incorporación basándose en una característica de la etiqueta, tal como fluorescencia de la etiqueta; una característica del monómero de nucleótido tal como peso molecular o carga; un subproducto de la incorporación del nucleótido, tal como liberación de pirofosfato; o similares. En las realizaciones, en las que dos o más nucleótidos diferentes están presentes en un reactivo de secuenciación, los nucleótidos diferentes pueden ser distinguibles entre sí, o como alternativa, las dos o más etiquetas diferentes pueden ser las indistinguibles en las técnicas de detección que se usan. Por ejemplo, los diferentes nucleótidos presentes en un reactivo de secuenciación pueden tener diferentes etiquetas y pueden distinguirse usando ópticas apropiadas como se ilustra por los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

Realizaciones preferidas incluyen técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPi) a medida que se incorporan nucleótidos particulares en la cadena naciente (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. y Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release". *Analytical Biochemistry* 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing". *Genome Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. y Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate". *Science* 281(5375), 363; Patente de Estados Unidos N.º 6.210.891; Patente de Estados Unidos N.º 6.258.568 y Patente de Estados Unidos N.º 6.274.320). En la pirosecuenciación, el PPi liberado puede detectarse convirtiéndose inmediatamente en trifosfato de adenosina (ATP) mediante ATP sulfurilasa, y el nivel de ATP generado se detecta a través de fotones producidos por luciferasa. Los ácidos nucleicos que hay que secuenciar pueden unirse a características en una matriz y la matriz puede representarse para capturar las señales quimioluminiscentes que se producen debido a la incorporación de un nucleótido a las características de la matriz. Puede obtenerse una imagen después de que la matriz se trata con un tipo de nucleótido particular (por ejemplo, A, T, C o G). Las imágenes obtenidas después de la adición de cada tipo de nucleótido diferirán con respecto a qué características se detectan en la matriz. Estas diferencias en la imagen reflejan el diferente contenido de secuencia de las características en la matriz. Sin embargo, las ubicaciones relativas de cada característica permanecerán sin

cambios en las imágenes. Las imágenes pueden almacenarse, procesarse y analizarse usando los métodos expuestos en este documento. Por ejemplo, imágenes obtenidas después del tratamiento de la matriz con cada diferente tipo de nucleótido pueden tratarse de la misma manera que se ilustra en este documento para imágenes obtenidas de diferentes canales de detección para métodos de secuenciación basada en terminadores reversibles.

En otro tipo ilustrativo de SBS, se consigue secuenciación de ciclo mediante adición escalonada de nucleótidos de terminación reversible que contienen, por ejemplo, una etiqueta de colorante escindible o fotoblanqueable como se describe, por ejemplo, en el documento WO 04/018497 y la Patente de Estados Unidos N.º 7.057.026. Este enfoque se está comercializando por Solexa (actualmente Illumina Inc.), y se describe también en el documento WO 91/06678 y el documento WO 07/123.744. La disponibilidad de terminadores etiquetados fluorescentemente en la que tanto la terminación puede reservarse como la etiqueta fluorescente escindible facilita la secuenciación de terminación reversible cíclica (CRT). También pueden diseñarse polimerasas para incorporar de forma eficiente y extender a partir de estos nucleótidos modificados.

Preferentemente en realizaciones de secuenciación basada en terminadores reversibles, las etiquetas sustancialmente no inhiben la extensión en condiciones de reacciones de SBS. Sin embargo, las etiquetas de detección pueden ser extraíbles, por ejemplo, por escisión o degradación. Pueden capturarse imágenes después de la incorporación de etiquetas en características de ácido nucleico en matriz. En realizaciones particulares, cada ciclo implica el suministro simultáneo de cuatro tipos de nucleótidos diferentes a la matriz y cada tipo de nucleótido tiene una etiqueta espectralmente distinta. A continuación pueden obtenerse cuatro imágenes, usando cada una un canal de detección que es selectivo para una de las cuatro etiquetas diferentes. Como alternativa, pueden añadirse secuencialmente diferentes tipos de nucleótidos y puede obtenerse una imagen de la matriz entre cada etapa de adición. En tales realizaciones cada imagen mostrará características de ácido nucleico que han incorporado nucleótidos de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica. Sin embargo, la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de tales métodos de SBS de terminadores reversibles pueden almacenarse, procesarse y analizarse como se expone en este documento. Después de la etapa de captura de imagen, pueden eliminarse etiquetas y pueden eliminarse fracciones de terminación reversibles para ciclos posteriores de adición y detección de nucleótidos. La eliminación de las etiquetas después de que se hayan detectado en un ciclo particular y antes de un ciclo posterior puede proporcionar la ventaja de reducir señal de fondo y diafonía entre ciclos. A continuación se exponen ejemplos de etiquetas útiles y métodos de eliminación.

En realizaciones particulares algunos o todos los monómeros de nucleótido pueden incluir terminadores reversibles. En tales realizaciones, los terminadores reversibles/flúores escindibles pueden incluir flúor enlazado a la fracción de ribosa a través de un enlace 3' éster (Metzker, Genome Res. 15:1767-1776 (2005)). Otros enfoques han separado la química de terminación de la escisión de la etiqueta de fluorescencia (Ruparel et al., Proc Natl Acad Sci USA 102: 5932-7 (2005)). Ruparel et al describieron el desarrollo de terminadores reversibles que usaban un pequeño grupo 3' alilo para bloquear la extensión, pero podría desbloquearse fácilmente mediante un tratamiento corto con un catalizador de paladio. El fluoróforo se unió a la base a través de un enlazante fotoescindible que podría escindirse fácilmente por una exposición de 30 segundos a luz UV de longitud de onda larga. Por lo tanto, puede usarse o bien reducción de disulfuro o bien fotoescisión como un enlazante escindible. Otro enfoque a terminación reversible es el uso de terminación natural que garantiza después de la colocación de un colorante voluminoso en un dNTP. La presencia de un colorante voluminoso cargado en el dNTP puede actuar como un terminador efectivo a través de impedimento estérico y/o electroestático. La presencia de un evento de incorporación evita incorporaciones adicionales a no ser que se elimine el colorante. La escisión del colorante elimina el flúor e invierte eficazmente la terminación. Ejemplos de nucleótidos modificados también se describen en la Patente de Estados Unidos N.º 7.427.673 y la Patente de Estados Unidos N.º 7.057.026.

Sistemas y métodos de SBS ilustrativos adicionales que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2007/0166705, Publicación de solicitud de Patente de Estados Unidos N.º 2006/0188901, Patente de Estados Unidos N.º 7.057.026, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0240439, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0281109, Publicación PCT N.º WO 05/065814, Publicación de Solicitud de Patente de Estados Unidos N.º 2005/0100900, Publicación PCT N.º WO 06/064199, Publicación PCT N.º WO 07/010.251, Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0270305 y Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0260372.

Algunas realizaciones pueden utilizar detección de cuatro diferentes nucleótidos usando menos de cuatro etiquetas diferentes. Por ejemplo, puede realizarse SBS utilizando métodos y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, pueden detectarse un par de tipos de nucleótidos en la misma longitud de onda, pero distinguirse basándose en una diferencia en intensidad para un miembro del par en comparación con el otro, o basándose en un cambio a un miembro del par (por ejemplo, a través de modificación química, modificación fotoquímica o modificación física) que provoca que la señal aparente aparezca o desaparezca en comparación con la señal detectada para el otro miembro del par. Como un segundo ejemplo, pueden detectarse tres de cuatro diferentes tipos de nucleótidos en condiciones particulares mientras un cuarto tipo

de nucleótido carece de una etiqueta que es detectable en esas condiciones, o se detecta mínimamente en esas condiciones (por ejemplo, detección mínima debido a fluorescencia de fondo, etc.). La incorporación de los primeros tres tipos de nucleótidos en un ácido nucleico puede determinarse basándose en la presencia de sus respectivas señales y la incorporación del cuarto tipo de nucleótido en el ácido nucleico puede determinarse basándose en la ausencia o detección mínima de cualquier señal. Como un tercer ejemplo, un tipo de nucleótido puede incluir etiqueta o etiquetas que se detectan en dos canales diferentes, mientras que otros tipos de nucleótidos se detectan en no más de uno de los canales. Las tres configuraciones ilustrativas anteriormente mencionadas no se consideran mutuamente exclusivas y pueden usarse en diversas combinaciones. Una realización ilustrativa que combina los tres ejemplos, es un método de SBS basado en fluorescencia que usa un primer tipo de nucleótido que se detecta en un primer canal (por ejemplo, dATP que tiene una etiqueta que se detecta en el primer canal cuando se excita por una primera longitud de onda de excitación), un segundo tipo de nucleótido que se detecta en un segundo canal (por ejemplo, dCTP que tiene una etiqueta que se detecta en el segundo canal cuando se excita por una segunda longitud de onda de excitación), un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal (por ejemplo, dTTP que tiene al menos una etiqueta que se detecta en ambos canales cuando se excita por la primera y/o segunda longitud de onda de excitación) y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o se detecta mínimamente, en cada canal (por ejemplo, dGTP que no tiene etiqueta).

Además, como se describe en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232, pueden obtenerse datos de secuenciación usando un único canal. En tales así llamados enfoques de secuenciación de un colorante, el primer tipo de nucleótido se etiqueta pero la etiqueta se elimina después de que se genera la primera imagen, y el segundo tipo de nucleótido se etiqueta únicamente después de que se genera una primera imagen. El tercer tipo de nucleótido retiene su etiqueta tanto en la primera como segunda imágenes, y el cuarto tipo de nucleótido permanece sin etiquetar en ambas imágenes.

Algunas realizaciones pueden utilizar secuenciación mediante técnicas de ligado. Tales técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de tales oligonucleótidos. Los oligonucleótidos habitualmente tienen diferentes etiquetas que se correlacionan con la identidad de un nucleótido particular en una secuencia a la que los oligonucleótidos hibridan. Como con otros métodos de SBS, las imágenes pueden obtenerse después del tratamiento de una matriz de características de ácido nucleico con los reactivos de secuenciación etiquetados. Cada imagen mostrará características de ácido nucleico que han incorporado etiquetas de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica, pero la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de métodos de secuenciación basados en ligado pueden almacenarse, procesarse y analizarse como se expone en este documento. Sistemas y métodos de SBS ilustrativos que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Patente de Estados Unidos N.º 6.969.488, Patente de Estados Unidos N.º 6.172.218, y Patente de Estados Unidos N.º 6.306.597.

Algunas realizaciones pueden utilizar secuenciación por nanoporos (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing". Trends Biotechnol. 18, 147-151 (2000); Deamer, D. Y D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, y J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). En tales realizaciones, el ácido nucleico objetivo pasa a través de un nanoporo. El nanoporo puede ser un poro sintético o proteína de membrana biológica, tal como α -hemolisina. Como el ácido nucleico objetivo pasa a través del nanoporo, cada par de bases puede identificarse midiendo fluctuaciones en la conductancia eléctrica del poro. (Patente de Estados Unidos N.º 7.001.792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores". Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis". Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution". J. Am. Chem. Soc. 130, 818-820 (2008)). Pueden almacenarse, procesarse y analizarse datos obtenidos de secuenciación por nanoporos como se expone en este documento. En particular, los datos pueden tratarse como una imagen de acuerdo con el tratamiento ilustrativo de imágenes ópticas y otras imágenes que se expone en este documento.

Algunas realizaciones pueden utilizar métodos que implican la supervisión en tiempo real de actividad de ADN polimerasa. Pueden detectarse incorporaciones de nucleótidos a través de interacciones de transferencia de energía de resonancia de fluorescencia (FRET) entre una polimerasa que porta fluoróforo y nucleótidos etiquetados con γ -fosfato como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.329.492 y Patente de Estados Unidos N.º 7.211.414 o pueden detectarse incorporaciones de nucleótidos con guías de ondas en modo cero como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.315.019 y usando análogos de nucleótidos fluorescentes y polimerasas diseñadas como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.405.281 y Publicación de Solicitud de Patente de Estados Unidos N.º 2008/0108082. La iluminación puede restringirse a un volumen de escala de zeptolitro alrededor de una polimerasa anclada a superficie de tal forma que la incorporación de nucleótidos etiquetados fluorescentemente puede observarse con poco fondo (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations". Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time". Opt. Lett. 33, 1026-1028 (2008); Koriach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures". Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). Las imágenes obtenidas a partir de tales métodos pueden

almacenarse, procesarse y analizarse como se expone en este documento.

Los métodos de SBS anteriores pueden efectuarse ventajosamente en formatos de multiplex de tal forma que se manipulan simultáneamente múltiples ácidos nucleicos objetivo diferentes. En realizaciones particulares, diferentes ácidos nucleicos objetivo pueden tratarse en un recipiente de reacción común o una superficie de un sustrato particular. Esto permite el suministro de reactivos de secuenciación, eliminación de reactivos sin reaccionar y detección de eventos de incorporación de una manera multiplex. En las realizaciones que usan ácidos nucleicos objetivo enlazados a superficie, los ácidos nucleicos objetivo pueden ser en un formato de matriz. En un formato de matriz, los ácidos nucleicos objetivo pueden enlazarse habitualmente a una superficie de una manera espacialmente distinguible. Los ácidos nucleicos objetivo pueden enlazarse por unión covalente directa, unión a una perla u otra partícula o enlace a una polimerasa u otra molécula que se une a la superficie. La matriz puede incluir una única copia de un ácido nucleico objetivo en cada sitio (también denominado como una característica) o múltiples copias que tiene la misma secuencia pueden estar presentes en cada sitio o característica. Múltiples copias pueden producirse mediante métodos de amplificación tales como, amplificación de puente o PCR en emulsión como se describe en detalle adicional a continuación.

Los métodos expuestos en este documento pueden usar matrices que tienen características en cualquiera de diversas densidades incluyendo, por ejemplo, al menos aproximadamente 10 características/cm², 100 características/cm², 500 características/cm², 1.000 características/cm², 5.000 características/cm², 10.000 características/cm², 50.000 características/cm², 100.000 características/cm², 1.000.000 características/cm², 5.000.000 características/cm², o mayores.

Sistemas

Un sistema capaz de efectuar un método expuesto en este documento, si se integra con capacidades de detección o no, puede incluir un controlador de sistema que es capaz de ejecutar un conjunto de instrucciones para realizar una o más etapas de un método, técnica o proceso expuesto en este documento. Por ejemplo, las instrucciones pueden dirigir la realización de etapas para crear un conjunto de amplicones *in situ*. Opcionalmente, las instrucciones pueden dirigir adicionalmente la realización de etapas para detectar ácidos nucleicos usando métodos expuestos anteriormente en este documento. Un controlador de sistema útil puede incluir cualquier sistema basado en procesador o basado en microprocesador, incluyendo sistemas que usan microcontroladores, ordenadores de conjunto reducido de instrucciones (RISC), circuitos integrados específicos de la aplicación (ASIC), campos de matriz de puertas programables (FPGA), circuitos lógicos y cualquier otro circuito o procesador capaz de ejecutar funciones descritas en este documento. Un conjunto de instrucciones para un controlador de sistema puede ser en forma de un programa de software. Como se usa en este documento, los términos "software" y "firmware" son intercambiables, e incluyen cualquier programa informático almacenado en memoria para su ejecución por un ordenador, incluyendo memoria RAM, memoria ROM, memoria EPROM, memoria EEPROM y memoria RAM no volátil (NVRAM). El software puede ser en diversas formas tales como software de sistema o software de aplicación. Además, el software puede ser en forma de una colección de programas separados, o un módulo de programa dentro de un programa mayor o una porción de un módulo de programa. El software también puede incluir programación modular en forma de programación orientada a objeto.

Se apreciará que cualquiera de los procesos de secuenciación anteriormente descritos puede incorporarse en los métodos y/o sistemas descritos en este documento. Adicionalmente, se apreciará que pueden implementarse fácilmente otros procesos de secuenciación conocidos para su uso con los métodos y/o sistemas descritos en este documento. Se apreciará también que los métodos y sistemas descritos en este documento se diseñan para ser aplicables con cualquier tecnología de secuenciación de ácido nucleico. Adicionalmente, se apreciará que los métodos y sistemas descritos en este documento tiene incluso una aplicabilidad más amplia para cualquier campo en el que el rastreo y análisis de características en un espécimen con el paso del tiempo o desde diferentes perspectivas es importante. Por ejemplo, los métodos y sistemas descritos en este documento pueden aplicarse donde datos de imagen obtenidos por tecnologías de formación de imágenes de vigilancia, aéreas o por satélite y similares se adquieren en puntos de tiempo o perspectivas diferentes y analizan.

Ejemplos

EJEMPLO 1

CORRECCIÓN DE AJUSTE DE FASE EMPÍRICA EN DATOS DE 2 CANALES

El ajuste de fase empírico se implementó en un sistema de secuenciación de 2 canales ejecutando una secuenciación de genoma completa de muestras humanas. La Figura 1 muestra datos representativos de una placa particular y un ciclo particular. Específicamente, como se muestra en la Figura 1B, usando el método de corrección de ajuste de fase descrito a continuación, una resolución dramáticamente aumentada resulta para intensidades asignadas a cada una de las cuatro bases.

La idea fundamental del algoritmo de corrección empírica es que la corrección de ajuste de fase maximiza la

depuración acumulada de los datos. Usando el algoritmo de corrección descrito anteriormente, es posible iterar sobre todos los valores de corrección de ajuste de fase y establecer cuál proporciona los mejores resultados. Un ejemplo se expone en la Figura 2, que representa datos de intensidad para un sistema de dos canales que se ha sometido a diversas correcciones de ajuste de fase. En la izquierda está el ciclo 150 de la ejecución de secuenciación, en el que el ajuste de fase está infracorrecto. En el centro están los datos óptimamente corregidos. En la derecha están los datos sobrecorrectos. Claramente, la depuración media de los datos se maximiza cuando la tasa de ajuste de fase supuesta es el valor verdadero.

Este conocimiento puede aprovecharse para estimar un parámetro de corrección de ajuste y preajuste de fase en cada ciclo que maximiza la depuración para ese ciclo. Para lograr esto, se implementa una corrección de ajuste de fase de primer orden:

$$I(\text{ciclo}) = I(\text{ciclo}) A * I(\text{ciclo}-1) B * I(\text{ciclo}+1)$$

Normalmente, las constantes A y B se calculan a partir de las estimadas tasas de ajuste/preajuste de fase y se ponderan por el número de ciclo. En una realización que usa corrección de ajuste de fase empírica, el método puede optimizar sobre A y B en cada ciclo usando una búsqueda de patrones. La función de coste es el número de grupos que no pasan un filtro de depuración. Por lo tanto, se seleccionan A y B para maximizar la calidad de datos.

Para optimizar el coste de cálculo de corregir de forma efectiva en muchas tasas de ajuste de fase diferentes, a continuación eligiendo la mejor, los valores A y B óptimos en cada ciclo se guardaron en el siguiente archivo: \Data\Intensities\BaseCalls\Phasing\EmpiricalPhasingCorrection_lane_read_tile.txt.

Estos archivos de datos tienen la siguiente estructura:

Cycle (Ciclo) PhasingCorrection (Corrección de ajuste de fase) PrephasingCorrection (Corrección de preajuste de fase)

Para determinar la tasa de ajuste o preajuste de fase, la lista de corrección de ajuste de fase se representó gráficamente por ciclo. La tasa de ajuste de fase es la pendiente de la línea resultante.

EJEMPLO 2

CORRECCIÓN DE AJUSTE DE FASE EMPÍRICA EN DATOS DE 4 CANALES DE DIVERSIDAD BAJA

Secuenciación de cuatro canales de muestras de diversidad baja tal como amplicones individuales presenta varios retos, incluyendo bajo rendimiento, bajo %PF, y puntuaciones de calidad bajas. Incluso cuando un genoma del bacteriófago (PhiX) conocido se introdujo en la muestra hasta niveles aproximándose al 50 %, estos retos persisten.

Una ejecución de secuenciación de un solo amplicón se realizó utilizando corrección de ajuste de fase para proporcionar datos de calidad alta en condiciones de diversidad extremadamente baja. En este experimento, se realizaron 3 ejecuciones de un solo amplicón separadas con ejecuciones de extremos emparejados de 101 ciclos de cada extremo. Se usó una versión de software de análisis en tiempo real (RTA versión 1.17.23) para analizar los datos de cuatro canales. Esta versión de RTA incluía ajuste de fase empírico. En todos los experimentos, todas las densidades de grupos fueron mayores de 1000k/mm² y el número de grupos que pasaron filtro fue mayor del 90 %.

Todos los datos de secuenciación tenían una puntuación de calidad porcentual superior a Q30 del 93 %. Estos resultados demuestran que el ajuste de fase empírico en datos de secuenciación de diversidad baja produce una calidad de datos superior.

El término comprendiendo/que comprende se concibe en este documento para ser abierto, incluyendo no únicamente los elementos citados, sino incluyendo adicionalmente cualquier elemento adicional.

Un número de las realizaciones se han descrito. Sin embargo, se entenderá que pueden hacerse diversas modificaciones.

La siguiente descripción es con respecto a las Figuras 3-5. Realizaciones descritas en lo sucesivo también se describen en la Solicitud Provisional de Estados Unidos N.º 61/915.426, presentada el jueves, 12 de diciembre de 2013.

El análisis de datos de imagen presenta un número de retos, especialmente con respecto a la comparación de imágenes de un artículo o estructura que se capturan desde diferentes puntos de referencia. La mayoría de metodología de análisis de imágenes emplea, al menos en parte, etapas para alinear múltiples imágenes separadas entre sí basándose en características o elementos presentes en ambas imágenes. Diversas realizaciones de las composiciones y métodos divulgados en este documento mejoran métodos anteriores para el análisis de imágenes. Algunos métodos anteriores para análisis de imágenes se exponen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537 presentada en 13 de enero de 2011 y titulada, "DATA PROCESSING SYSTEM AND METHODS".

En la actualidad, se han desarrollado herramientas que adquieren y analizan datos de imagen generados en puntos de tiempo o perspectivas diferentes. Algunos ejemplos incluyen herramientas para análisis de imágenes de satélite y herramientas de biología molecular para secuenciar y caracterizar la identidad molecular de un espécimen. En cualquier tal sistema, adquirir y almacenar grandes números de imágenes de alta calidad habitualmente requiere cantidades masivas de capacidad de almacenamiento. Adicionalmente, una vez adquiridos y almacenados, el análisis de datos de imagen puede volverse intensivo en recursos y puede interferir con capacidad de procesamiento de otras funciones, tales como adquisición constante y almacenamiento de datos de imagen adicionales. Como tal, serían beneficiosos métodos y sistemas que mejoren la velocidad y precisión de análisis de la adquisición y análisis de datos de imagen.

En el campo de la biología molecular, uno de los procesos para secuenciación de ácido nucleico en uso es secuenciación por síntesis. La técnica puede aplicarse a proyectos de secuenciación masivamente paralelos. Por ejemplo, usando una plataforma automatizada, es posible efectuar cientos de miles de reacciones de secuenciaciones simultáneamente. Por lo tanto, la divulgación se refiere a instrumentos y métodos para adquirir, almacenar y analizar datos de imagen generados durante secuenciación de ácido nucleico.

Enormes ganancias en la cantidad de datos que pueden adquirirse y almacenarse hacen incluso más beneficiosos a los métodos de análisis de imagen simplificados. Por ejemplo, los métodos de análisis de imagen descritos en este documento permiten que tanto diseñadores como usuarios finales hagan un uso eficiente de hardware informático existente. Por consiguiente, en este documento se presentan métodos y sistemas que reducen la carga de cálculo del procesamiento de datos ante el rápido crecimiento de la emisión de datos. Por ejemplo, en el campo de secuenciación de ADN, los rendimientos han escalado 15 veces durante el transcurso del último año, y ahora puede alcanzar cientos de gigabases en una única ejecución de un dispositivo de secuenciación de ADN. Si los requisitos de infraestructura de cálculo crecen exponencialmente, los experimentos de genoma a gran escala permanecerían fuera del alcance de la mayoría de investigadores. Por lo tanto, la generación de más datos de secuencia sin procesar aumentará la necesidad de análisis secundario y almacenamiento de datos, haciendo extremadamente valiosa la optimización de transporte y almacenamiento de datos. Algunas realizaciones de los métodos y sistemas presentados en este documento pueden reducir el tiempo, hardware, interconexión de redes y requisitos de infraestructura de laboratorio necesarios para producir datos de secuencia usables.

Como se usa en este documento, una "característica" es un área de interés dentro de un espécimen o campo de visión. Cuando se usa en conexión con dispositivos de micromatriz u otros dispositivos de analítica molecular, una característica se refiere al área ocupada por moléculas similares o idénticas. Por ejemplo, una característica puede ser un oligonucleótido amplificado o cualquier otro grupo de un polinucleótido o polipéptido con una misma o similar secuencia. En otras realizaciones, una característica puede ser cualquier elemento o grupo de elementos que ocupa un área física en un espécimen. Por ejemplo, una característica podría ser una parcela de tierra, una masa de agua o similar. Cuando se representa una característica, cada característica tendrá alguna área. Por lo tanto, en muchas realizaciones, una característica no es meramente un píxel.

Las distancias entre características pueden describirse de cualquier número de formas. En algunas realizaciones, las distancias entre características pueden describirse a partir del centro de una característica hasta el centro de otra característica. En otras realizaciones, las distancias pueden describirse desde el borde de una característica hasta el borde de otra característica, o entre los puntos identificables más externos de cada característica. El borde de una característica puede describirse como el límite físico teórico o real en un chip, o algún punto dentro del límite de la característica. En otras realizaciones, las distancias pueden describirse en relación con un punto fijo en el espécimen o en la imagen del espécimen.

Pueden secuenciarse múltiples copias de ácidos nucleicos en una característica, por ejemplo, proporcionando una base de nucleótido etiquetada a la matriz de moléculas, extendiendo de este modo un cebador hibridado a un ácido nucleico dentro de una característica para producir una señal que corresponde a una característica que comprende el ácido nucleico. En realizaciones preferidas, los ácidos nucleicos dentro de una característica son idénticos o sustancialmente idénticos entre sí.

En algunos de los métodos de análisis de imagen descritos en este documento, cada imagen en el conjunto de imágenes incluye señales de colores, en el que un color diferente corresponde a una base de nucleótido diferente. En algunos aspectos, cada imagen del conjunto de imágenes comprende señales que tienen un único color seleccionado de al menos cuatro colores diferentes. En ciertos aspectos, cada imagen en el conjunto de imágenes comprende señales que tienen un único color seleccionado de cuatro colores diferentes.

Con respecto a ciertos métodos de cuatro canales descritos en este documento, pueden secuenciarse ácidos nucleicos proporcionando, cuatro diferentes bases de nucleótidos etiquetadas a la matriz de moléculas para producir cuatro imágenes diferentes, comprendiendo cada imagen señales que tienen un único color, en el que el color de señal es diferente para cada una de las cuatro imágenes diferentes, produciendo de este modo un ciclo de cuatro imágenes de color que corresponde a los cuatro posibles nucleótidos presentes en una posición particular en el ácido nucleico. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de

color.

Con respecto a ciertos métodos de dos canales descritos en este documento, pueden secuenciarse ácidos nucleicos utilizando métodos y sistemas descritos en la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, puede secuenciarse un ácido nucleico proporcionando un primer tipo de nucleótido que se detecta en un primer canal, un segundo tipo de nucleótido que se detecta en un segundo canal, un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o está mínimamente, en ningún canal. En ciertos aspectos, tales métodos pueden comprender adicionalmente proporcionar bases de nucleótidos etiquetadas adicionales a la matriz de moléculas, produciendo de este modo una pluralidad de ciclos de imágenes de color.

Identificación de bases

En este documento se presentan métodos y sistemas para identificar una base de nucleótido en una secuencia de ácido nucleico o "identificación de bases". Identificación de bases se refiere al proceso de determinar una identificación de bases (A, C, G, T) para cada característica de una placa dada en un ciclo específico. Como un ejemplo, puede realizarse SBS utilizando métodos de dos canales y sistemas descritos en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Por ejemplo, en las realizaciones que hacen uso de detección de dos canales, se realiza identificación de bases extrayendo datos de imagen de dos imágenes, en lugar de cuatro. Debido a las diferencias fundamentales implicadas en identificación de bases de dos canales, enfoques tradicionales de identificación de bases como se aplican a detección de cuatro canales no es compatible con datos de dos canales. En vista de estas diferencias, se requiere un nuevo enfoque para identificación de bases. Por consiguiente, en este documento se presentan métodos y sistemas para identificación de bases en un sistema de 2 canales. En algunas realizaciones, los métodos comprenden ajustar iterativamente cuatro distribuciones gaussianas a datos de intensidad de dos canales. Cuando señales del canal 1 se representan gráficamente contra señales del canal 2, intensidad de señal habitualmente se segrega en cuatro poblaciones generales de intensidad. Como se muestra en la Figura 3, los datos desde un sistema de secuenciación de 2 canales pueden representarse gráficamente como valores de intensidad del canal 1 (eje x) frente a valores de intensidad del canal 2 (eje y). En realizaciones típicas, uno de los cuatro nucleótidos no tiene etiqueta (oscuro), tal como el nucleótido "G" mostrado en la Figura 3, que tiene una señal casi cero tanto en el canal 1 como el canal 2. Las señales desde una cierta porción de los puntos de datos se agrupan cerca del punto cero en cada eje. Análogamente, las señales desde una cierta porción de los puntos de datos etiquetados con una o ambas etiquetas (mostrados como nucleótidos "C", "A" y "T" en la Figura 3) forman poblaciones identificables cuando se representan gráficamente en un gráfico bidimensional tal como el mostrado en la Figura 3. Por lo tanto, por ejemplo, a diferencia de datos de secuenciación de cuatro canales, la propia intensidad de una etiqueta particular no codifica la base. En su lugar, la combinación de intensidades, [encendida, apagada], [apagada, encendida], [encendida, encendida], [apagada, apagada], proporciona la información de codificación para la identidad de base.

Los métodos y sistemas presentados en este documento proporcionan una herramienta para identificar la base asociada con un cualquier punto de datos particular en tales conjuntos de datos. Un objetivo de los métodos y sistemas presentados en este documento es separar las cuatro poblaciones de forma tan precisa como sea posible.

Clasificadores

En algunas realizaciones presentadas en este documento, se realiza identificación de bases ajustando un modelo matemático a un conjunto de datos de intensidad. Cualquier modelo matemático adecuado puede usarse en los métodos presentados en este documento para ajustar los datos de intensidad a un conjunto de distribuciones. Modelos matemáticos que pueden usarse en los métodos presentados en este documento pueden incluir clasificadores tales como, por ejemplo, un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, maximización de la expectativa, método basado en histogramas y similares.

Por ejemplo, en ciertas realizaciones, una o más distribuciones gaussianas se ajustan a un conjunto de datos de intensidad. En ciertas realizaciones, 4 distribuciones gaussianas se ajustan a un conjunto de datos de intensidad de dos canales, de tal forma que se aplica una distribución para cada uno de los cuatro nucleótidos representados en el conjunto de datos. En ciertas realizaciones, valores de intensidad pueden normalizarse antes del ajuste de una distribución gaussiana. Por ejemplo, como se muestra en la realización ilustrativa representada por la Figura 4, se normalizan valores de intensidad de modo que los percentiles 5 y 95 tienen valores de 0 y 1, respectivamente. A continuación se ajustan cuatro distribuciones gaussianas a los datos usando un algoritmo tal como, por ejemplo, un algoritmo de agrupamiento de maximización de la expectativa (EM). Los algoritmos EM se conocen en la técnica y son herramientas útiles para construir modelos estadísticos de la fuente de datos subyacente y generalizar naturalmente las bases de datos de grupos que contienen tanto datos de valores discretos como valores continuos. Por lo tanto, por ejemplo, en ciertas realizaciones, se aplica un algoritmo EM para maximizar iterativamente la probabilidad de observar los datos dados. Por ejemplo, un algoritmo EM se aplica para maximizar iterativamente esta probabilidad sobre la media y covarianza para cada una de las distribuciones gaussianas. En ciertas realizaciones, un subconjunto de los puntos de datos en un conjunto de datos se incluye en el cálculo. Adicionalmente o como alternativa, en ciertas realizaciones, todos o sustancialmente todos los puntos de datos en el conjunto de datos se incluyen en el cálculo.

Como resultado del algoritmo EM, para cada valor de X, Y (haciendo referencia a cada una de las dos intensidades de canal respectivamente) puede generarse un valor que representa la probabilidad de que un cierto valor de intensidad de X, Y pertenezca a una de las cuatro distribuciones. En una realización en la que cuatro bases proporcionan cuatro distribuciones separadas, cada valor de intensidad de X, Y también tendrá cuatro valores de probabilidad asociados, uno para cada una de las cuatro bases. El máximo de los cuatro valores de probabilidad indica la identificación de bases. Por lo tanto, como se muestra en la realización ilustrativa representada por la Figura 5, valores de intensidad para un conjunto de datos de dos canales se asignan a una identificación de bases después de realizar un ajuste gaussiano al conjunto de datos. Cada punto de datos en los gráficos en la Figura 5 tiene un color asociado con la identificación de bases asignada, que representa el máximo de los valores de predicción de probabilidad. Una comparación de los datos de identificación de bases mostrados en los dos gráficos en la Figura 5 indica que los métodos de identificación de bases presentados en este documento son altamente precisos y son robustos a diversos tipos de química de secuenciación. Por ejemplo, el panel izquierdo de la Figura 5 es un ejemplo de química que forma cuatro distribuciones de intensidades que forman un cuadrado cuando los valores de intensidad se representan gráficamente. En contraste, la intensidad representada gráficamente en el panel derecho tiene cuatro distribuciones de intensidades que se encuentran dentro de un triángulo, basándose en las menores intensidades del nucleótido de etiqueta doble. En ambos tipos de química, los métodos de identificación de bases presentados en este documento proporcionan identificaciones de bases precisas.

En las realizaciones de los métodos presentados en este documento, una puntuación de calidad también puede generarse basándose en el enfoque de distribución gaussiana a identificación de bases. Por ejemplo, la distancia de un punto al centro de la distribución "identificada" proporciona una medida de la pureza de la identificación de bases. Específicamente, cuanto más cerca se encuentre el punto de datos del centro de la distribución para la base identificada, mayor será la probabilidad de que la identificación de bases sea precisa. Cualquier método adecuado para calcular y expresar la relación entre distancia al centro y probablemente pureza de la identificación de bases puede usarse en los métodos proporcionados en este documento. En algunas realizaciones, la calidad o pureza de la identificación de bases para un punto de datos dado puede expresarse como la distancia al centroide más cercano dividido por la suma de todas las distancias a cada uno de los otros tres centroides. En algunas realizaciones, la calidad o pureza de la identificación de bases para un punto de datos dado puede expresarse como la distancia al centroide más cercano dividido por la distancia al segundo centroide más cercano, como se describe a continuación con respecto a filtrado de depuración.

Filtrado de depuración

También se presentan en este documento métodos de filtrado de grupos que tienen mala calidad. El término filtrado como se usa en relación con grupos e identificación de bases se refiere a descartar o ignorar el grupo como un punto de datos. Por lo tanto, cualquier grupo de mala intensidad o calidad puede filtrarse y no se incluyen en un conjunto de datos de salida. En ciertas realizaciones, la calidad de grupo se determina mediante una métrica denominada depuración. La depuración para identificación de bases de dos canales asumen un significado separado del uso del término en identificación de bases de cuatro canales. Por ejemplo, como se describe en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0020537, la depuración se define en términos de intensidad de un grupo ("zona") en relación con una zona cercana), y puede calcularse como el mayor valor de intensidad dividido por la suma del mayor valor de intensidad y el segundo mayor valor de intensidad, en la que los valores de intensidad se obtienen a partir de cuatro canales de color. Sin embargo, debido a que la identificación de bases de dos canales habitualmente utiliza nucleótidos sin etiqueta que emiten una señal muy baja o no emiten ninguna señal, las determinaciones de depuración tradicionales no son adecuadas para la identificación de bases de dos canales.

Por lo tanto, algunas realizaciones de la presente divulgación se refieren a determinar depuración de un grupo como una función de distancias relativas a centroides gaussianos. En algunas realizaciones, los grupos que no están lo suficientemente cerca de un centroide gaussiano particular en un número dado de ciclos reciben un valor de depuración bajo y se filtran. Por ejemplo, en una realización específica, la depuración puede calcularse usando la expresión:

$$\text{depuración} = 1 - D1 / (D1 + D2),$$

en la que D1 es la distancia al centroide gaussiano más cercano y D2 es la distancia al siguiente centroide más cercano. Métodos de ajustar distribuciones gaussianas a un conjunto de datos de dos canales se describen anteriormente en este documento en la sección que describe métodos de identificación de bases.

En algunas realizaciones, el filtrado de grupos de depuración baja tiene lugar en uno más puntos discretos durante una ejecución de secuencia. En algunas realizaciones, se produce el filtrado durante la generación de molde. Como alternativa o adicionalmente, en algunas realizaciones, el filtrado se produce después de un ciclo predefinido. En ciertas realizaciones, el filtrado se produce en o después del ciclo 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, o después del ciclo 30 o posterior. En realizaciones típicas, el filtrado se produce en el ciclo 25, de tal forma que se filtran los grupos que no están suficientemente cerca de un centroide

gaussiano en los primeros 25 ciclos.

Métodos de secuenciación

- 5 Los métodos descritos en este documento pueden usarse en conjunto con diversas técnicas de secuenciación de ácido nucleico. Particularmente técnicas aplicables son en las que los ácidos nucleicos se unen a ubicaciones fijas en una matriz de tal forma que sus posiciones relativas no cambian y en las que la matriz se representa repetidamente. Son particularmente aplicables realizaciones en las que se obtienen imágenes en diferentes canales de color, por ejemplo, que coinciden con diferentes etiquetas usadas para distinguir un tipo de base de nucleótido de otro. En algunas realizaciones, el proceso para determinar la secuencia de nucleótido de un ácido nucleico objetivo puede ser un proceso automatizado. Realizaciones preferidas incluyen técnicas de secuenciación por síntesis ("SBS").

- 15 Las técnicas de SBS generalmente implican la extensión enzimática de una cadena de ácido nucleico naciente a través de la adición iterativa de nucleótidos contra una cadena molde. En métodos tradicionales de SBS, puede proporcionarse un monómero de un solo nucleótido a un nucleótido objetivo en la presencia de una polimerasa en cada suministro. Sin embargo, en los métodos descritos en este documento, puede proporcionarse más de un tipo de monómero de nucleótido a un ácido nucleico objetivo en la presencia de una polimerasa en un suministro.

- 20 SBS puede utilizar monómeros de nucleótido que tienen una fracción de terminación o los que carecen de cualquier fracción de terminación. Métodos que utilizan monómeros de nucleótido que carecen de terminadores incluyen, por ejemplo, pirosecuenciación y secuenciación usando nucleótidos etiquetados con γ -fosfato, como se expone en detalle adicional a continuación. En métodos que usan monómeros de nucleótido que carecen de terminadores, el número de nucleótidos añadidos en cada ciclo es generalmente variable y depende de la secuencia molde y el modo de suministro de nucleótido. Para técnicas de SBS que utilizan monómeros de nucleótido que tienen una fracción de terminación, el terminador puede ser irreversible de forma efectiva en las condiciones de secuenciación usadas como en el caso para la secuenciación de Sanger tradicional que utiliza didesoxinucleótidos, o el terminador puede ser reversible como es el caso para los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

- 30 Técnicas de SBS pueden utilizar monómeros de nucleótido que tienen una fracción de etiqueta o los que carecen de una fracción de etiqueta. Por consiguiente, pueden detectarse eventos de incorporación basándose en una característica de la etiqueta, tal como fluorescencia de la etiqueta; una característica del monómero de nucleótido tal como peso molecular o carga; un subproducto de la incorporación del nucleótido, tal como liberación de pirofosfato; o similares. En las realizaciones, en las que dos o más nucleótidos diferentes están presentes en un reactivo de secuenciación, los nucleótidos diferentes pueden ser distinguibles entre sí, o como alternativa, las dos o más etiquetas diferentes pueden ser las indistinguibles en las técnicas de detección que se usan. Por ejemplo, los diferentes nucleótidos presentes en un reactivo de secuenciación pueden tener diferentes etiquetas y pueden distinguirse usando ópticas apropiadas como se ilustra por los métodos de secuenciación desarrollados por Solexa (actualmente Illumina, Inc.).

- 40 Realizaciones preferidas incluyen técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPi) a medida que se incorporan nucleótidos particulares en la cadena naciente (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. y Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release". *Analytical Biochemistry* 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing". *Genome Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. y Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate". *Science* 281(5375), 363; Patente de Estados Unidos N.º 6.210.891; Patente de Estados Unidos N.º 6.258.568 y Patente de Estados Unidos N.º 6.274.320). En la pirosecuenciación, el PPi liberado puede detectarse convirtiéndose inmediatamente en trifosfato de adenosina (ATP) mediante ATP sulfúrilasa, y el nivel de ATP generado se detecta a través de fotones producidos por luciferasa. Los ácidos nucleicos que hay que secuenciar pueden unirse a características en una matriz y la matriz puede representarse para capturar las señales quimioluminiscentes que se producen debido a la incorporación de un nucleótido a las características de la matriz. Puede obtenerse una imagen después de que la matriz se trata con un tipo de nucleótido particular (por ejemplo, A, T, C o G). Las imágenes obtenidas después de la adición de cada tipo de nucleótido diferirán con respecto a qué características se detectan en la matriz. Estas diferencias en la imagen reflejan el diferente contenido de secuencia de las características en la matriz. Sin embargo, las ubicaciones relativas de cada característica permanecerán sin cambios en las imágenes. Las imágenes pueden almacenarse, procesarse y analizarse usando los métodos expuestos en este documento. Por ejemplo, imágenes obtenidas después del tratamiento de la matriz con cada diferente tipo de nucleótido pueden tratarse de la misma manera que se ilustra en este documento para imágenes obtenidas de diferentes canales de detección para métodos de secuenciación basada en terminadores reversibles.

- 60 En otro tipo ilustrativo de SBS, se consigue secuenciación de ciclo mediante adición escalonada de nucleótidos de terminación reversible que contienen, por ejemplo, una etiqueta de colorante escindible o fotoblanqueable como se describe, por ejemplo, en el documento WO 04/018497 y la Patente de Estados Unidos N.º 7.057.026. Este enfoque se está comercializando por Solexa (actualmente Illumina Inc.), y se describe también en el documento WO 91/06678 y el documento WO 07/123.744. La disponibilidad de terminadores etiquetados fluorescentemente en la que tanto la terminación puede reservarse como la etiqueta fluorescente escindible facilita la secuenciación de terminación reversible cíclica (CRT). También pueden diseñarse polimerasas para incorporar de forma eficiente y extender a partir

de estos nucleótidos modificados.

Preferentemente en realizaciones de secuenciación basada en terminadores reversibles, las etiquetas sustancialmente no inhiben la extensión en condiciones de reacciones de SBS. Sin embargo, las etiquetas de detección pueden ser extraíbles, por ejemplo, por escisión o degradación. Pueden capturarse imágenes después de la incorporación de etiquetas en características de ácido nucleico en matriz. En realizaciones particulares, cada ciclo implica el suministro simultáneo de cuatro tipos de nucleótidos diferentes a la matriz y cada tipo de nucleótido tiene una etiqueta espectralmente distinta. A continuación pueden obtenerse cuatro imágenes, usando cada una un canal de detección que es selectivo para una de las cuatro etiquetas diferentes. Como alternativa, pueden añadirse secuencialmente diferentes tipos de nucleótidos y puede obtenerse una imagen de la matriz entre cada etapa de adición. En tales realizaciones cada imagen mostrará características de ácido nucleico que han incorporado nucleótidos de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica. Sin embargo, la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de tales métodos de SBS de terminadores reversibles pueden almacenarse, procesarse y analizarse como se expone en este documento. Después de la etapa de captura de imagen, pueden eliminarse etiquetas y pueden eliminarse fracciones de terminación reversibles para ciclos posteriores de adición y detección de nucleótidos. La eliminación de las etiquetas después de que se hayan detectado en un ciclo particular y antes de un ciclo posterior puede proporcionar la ventaja de reducir señal de fondo y diafonía entre ciclos. A continuación se exponen ejemplos de etiquetas útiles y métodos de eliminación.

En realizaciones particulares algunos o todos los monómeros de nucleótido pueden incluir terminadores reversibles. En tales realizaciones, los terminadores reversibles/flúores escindibles pueden incluir flúor enlazado a la fracción de ribosa a través de un enlace 3' éster (Metzker, *Genome Res.* 15:1767-1776 (2005)). Otros enfoques han separado la química de terminación de la escisión de la etiqueta de fluorescencia (Ruparel et al., *Proc Natl Acad Sci USA* 102: 5932-7 (2005)). Ruparel et al describieron el desarrollo de terminadores reversibles que usaban un pequeño grupo 3' alilo para bloquear la extensión, pero podría desbloquearse fácilmente mediante un tratamiento corto con un catalizador de paladio. El fluoróforo se unió a la base a través de un enlazante fotoescindible que podría escindirse fácilmente por una exposición de 30 segundos a luz UV de longitud de onda larga. Por lo tanto, puede usarse o bien reducción de disulfuro o bien fotoescisión como un enlazante escindible. Otro enfoque a terminación reversible es el uso de terminación natural que garantiza después de la colocación de un colorante voluminoso en un dNTP. La presencia de un colorante voluminoso cargado en el dNTP puede actuar como un terminador efectivo a través de impedimento estérico y/o electroestático. La presencia de un evento de incorporación evita incorporaciones adicionales a no ser que se elimine el colorante. La escisión del colorante elimina el flúor e invierte eficazmente la terminación. Ejemplos de nucleótidos modificados también se describen en la Patente de Estados Unidos N.º 7.427.673 y la Patente de Estados Unidos N.º 7.057.026.

Sistemas y métodos de SBS ilustrativos adicionales que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Publicación de Solicitud de Patente de Estados Unidos N.º 2007/0166705, Publicación de solicitud de Patente de Estados Unidos N.º 2006/0188901, Patente de Estados Unidos N.º 7.057.026, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0240439, Publicación de Solicitud de Patente de Estados Unidos N.º 2006/0281109, Publicación PCT N.º WO 05/065814, Publicación de Solicitud de Patente de Estados Unidos N.º 2005/0100900, Publicación PCT N.º WO 06/064199, Publicación PCT N.º WO 07/010.251, Publicación de Solicitud de Patente de Estados Unidos N.º 2012/0270305 y Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0260372.

Algunas realizaciones pueden utilizar detección de cuatro diferentes nucleótidos usando menos de cuatro etiquetas diferentes. Por ejemplo, puede realizarse SBS utilizando métodos y sistemas descritos en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232. Como un primer ejemplo, pueden detectarse un par de tipos de nucleótidos en la misma longitud de onda, pero distinguirse basándose en una diferencia en intensidad para un miembro del par en comparación con el otro, o basándose en un cambio a un miembro del par (por ejemplo, a través de modificación química, modificación fotoquímica o modificación física) que provoca que la señal aparente aparezca o desaparezca en comparación con la señal detectada para el otro miembro del par. Como un segundo ejemplo, pueden detectarse tres de cuatro diferentes tipos de nucleótidos en condiciones particulares mientras un cuarto tipo de nucleótido carece de una etiqueta que es detectable en esas condiciones, o se detecta mínimamente en esas condiciones (por ejemplo, detección mínima debido a fluorescencia de fondo, etc.). La incorporación de los primeros tres tipos de nucleótidos en un ácido nucleico puede determinarse basándose en la presencia de sus respectivas señales y la incorporación del cuarto tipo de nucleótido en el ácido nucleico puede determinarse basándose en la ausencia o detección mínima de cualquier señal. Como un tercer ejemplo, un tipo de nucleótido puede incluir etiqueta o etiquetas que se detectan en dos canales diferentes, mientras que otros tipos de nucleótidos se detectan en no más de uno de los canales. Las tres configuraciones ilustrativas anteriormente mencionadas no se consideran mutuamente exclusivas y pueden usarse en diversas combinaciones. Una realización ilustrativa que combina los tres ejemplos, es un método de SBS basado en fluorescencia que usa un primer tipo de nucleótido que se detecta en un primer canal (por ejemplo, dATP que tiene una etiqueta que se detecta en el primer canal cuando se excita por una primera longitud de onda de excitación), un segundo tipo de nucleótido que se detecta en un segundo canal (por ejemplo, dCTP que tiene una etiqueta que se detecta en el segundo canal cuando se excita

por una segunda longitud de onda de excitación), un tercer tipo de nucleótido que se detecta tanto en el primer como el segundo canal (por ejemplo, dTTP que tiene al menos una etiqueta que se detecta en ambos canales cuando se excita por la primera y/o segunda longitud de onda de excitación) y un cuarto tipo de nucleótido que carece de una etiqueta que no se detecta, o se detecta mínimamente, en cada canal (por ejemplo, dGTP que no tiene etiqueta).

Además, como se describe en los materiales incorporados de la Publicación de Solicitud de Patente de Estados Unidos N.º 2013/0079232, pueden obtenerse datos de secuenciación usando un único canal. En tales así llamados enfoques de secuenciación de un colorante, el primer tipo de nucleótido se etiqueta pero la etiqueta se elimina después de que se genera la primera imagen, y el segundo tipo de nucleótido se etiqueta únicamente después de que se genera una primera imagen. El tercer tipo de nucleótido retiene su etiqueta tanto en la primera como segunda imágenes, y el cuarto tipo de nucleótido permanece sin etiquetar en ambas imágenes.

Algunas realizaciones pueden utilizar secuenciación mediante técnicas de ligado. Tales técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de tales oligonucleótidos. Los oligonucleótidos habitualmente tienen diferentes etiquetas que se correlacionan con la identidad de un nucleótido particular en una secuencia a la que los oligonucleótidos hibridan. Como con otros métodos de SBS, las imágenes pueden obtenerse después del tratamiento de una matriz de características de ácido nucleico con los reactivos de secuenciación etiquetados. Cada imagen mostrará características de ácido nucleico que han incorporado etiquetas de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al diferente contenido de secuencia de cada característica, pero la posición relativa de las características permanecerá sin cambios en las imágenes. Las imágenes obtenidas a partir de métodos de secuenciación basados en ligado pueden almacenarse, procesarse y analizarse como se expone en este documento. Sistemas y métodos de SBS ilustrativos que pueden utilizarse con los métodos y sistemas descritos en este documento se describen en la Patente de Estados Unidos N.º 6.969.488, Patente de Estados Unidos N.º 6.172.218, y Patente de Estados Unidos N.º 6.306.597.

Algunas realizaciones pueden utilizar secuenciación por nanoporos (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing". Trends Biotechnol. 18, 147-151 (2000); Deamer, D. Y D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, y J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003)). En tales realizaciones, el ácido nucleico objetivo pasa a través de un nanoporo. El nanoporo puede ser un poro sintético o proteína de membrana biológica, tal como α -hemolisina. Como el ácido nucleico objetivo pasa a través del nanoporo, cada par de bases puede identificarse midiendo fluctuaciones en la conductancia eléctrica del poro. (Patente de Estados Unidos N.º 7.001.792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores". Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis". Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution". J. Am. Chem. Soc. 130, 818-820 (2008)). Pueden almacenarse, procesarse y analizarse datos obtenidos de secuenciación por nanoporos como se expone en este documento. En particular, los datos pueden tratarse como una imagen de acuerdo con el tratamiento ilustrativo de imágenes ópticas y otras imágenes que se expone en este documento.

Algunas realizaciones pueden utilizar métodos que implican la supervisión en tiempo real de actividad de ADN polimerasa. Pueden detectarse incorporaciones de nucleótidos a través de interacciones de transferencia de energía de resonancia de fluorescencia (FRET) entre una polimerasa que porta fluoróforo y nucleótidos etiquetados con γ -fosfato como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.329.492 y Patente de Estados Unidos N.º 7.211.414 o pueden detectarse incorporaciones de nucleótidos con guías de ondas en modo cero como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.315.019 y usando análogos de nucleótidos fluorescentes y polimerasas diseñadas como se describe, por ejemplo, en la Patente de Estados Unidos N.º 7.405.281 y Publicación de Solicitud de Patente de Estados Unidos N.º 2008/0108082. La iluminación puede restringirse a un volumen de escala de zeptolitro alrededor de una polimerasa anclada a una superficie de tal forma que la incorporación de nucleótidos etiquetados fluorescentemente puede observarse con poco fondo (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations". Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time". Opt. Lett. 33, 1026-1028 (2008); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures". Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008)). Las imágenes obtenidas a partir de tales métodos pueden almacenarse, procesarse y analizarse como se expone en este documento.

Los métodos de SBS anteriores pueden efectuarse ventajosamente en formatos de multiplex de tal forma que se manipulan simultáneamente múltiples ácidos nucleicos objetivo diferentes. En realizaciones particulares, diferentes ácidos nucleicos objetivo pueden tratarse en un recipiente de reacción común o una superficie de un sustrato particular. Esto permite el suministro de reactivos de secuenciación, eliminación de reactivos sin reaccionar y detección de eventos de incorporación de una manera multiplex. En las realizaciones que usan ácidos nucleicos objetivo enlazados a superficie, los ácidos nucleicos objetivo pueden ser en un formato de matriz. En un formato de matriz, los ácidos nucleicos objetivo pueden enlazarse habitualmente a una superficie de una manera espacialmente distinguible. Los ácidos nucleicos objetivo pueden enlazarse por unión covalente directa, unión a una perla u otra partícula o enlace a una polimerasa u otra molécula que se une a la superficie. La matriz puede incluir una única copia de un ácido nucleico objetivo en cada sitio (también denominado como una característica) o múltiples copias que tiene la misma secuencia

pueden estar presentes en cada sitio o característica. Múltiples copias pueden producirse mediante métodos de amplificación tales como, amplificación de puente o PCR en emulsión como se describe en detalle adicional a continuación.

Los métodos expuestos en este documento pueden usar matrices que tienen características en cualquiera de diversas densidades incluyendo, por ejemplo, al menos aproximadamente 10 características/cm², 100 características/cm², 500 características/cm², 1.000 características/cm², 5.000 características/cm², 10.000 características/cm², 50.000 características/cm², 100.000 características/cm², 1.000.000 características/cm², 5.000.000 características/cm², o mayores.

Sistemas

Un sistema capaz de efectuar un método expuesto en este documento, si se integra con capacidades de detección o no, puede incluir un controlador de sistema que es capaz de ejecutar un conjunto de instrucciones para realizar una o más etapas de un método, técnica o proceso expuesto en este documento. Por ejemplo, las instrucciones pueden dirigir la realización de etapas para crear un conjunto de amplicones *in situ*. Opcionalmente, las instrucciones pueden dirigir adicionalmente la realización de etapas para detectar ácidos nucleicos usando métodos expuestos anteriormente en este documento. Un controlador de sistema útil puede incluir cualquier sistema basado en procesador o basado en microprocesador, incluyendo sistemas que usan microcontroladores, ordenadores de conjunto reducido de instrucciones (RISC), circuitos integrados específicos de la aplicación (ASIC), campos de matriz de puertas programables (FPGA), circuitos lógicos y cualquier otro circuito o procesador capaz de ejecutar funciones descritas en este documento. Un conjunto de instrucciones para un controlador de sistema puede ser en forma de un programa de software. Como se usa en este documento, los términos "software" y "firmware" son intercambiables, e incluyen cualquier programa informático almacenado en memoria para su ejecución por un ordenador, incluyendo memoria RAM, memoria ROM, memoria EPROM, memoria EEPROM y memoria RAM no volátil (NVRAM). El software puede ser en diversas formas tales como software de sistema o software de aplicación. Además, el software puede ser en forma de una colección de programas separados, o un módulo de programa dentro de un programa mayor o una porción de un módulo de programa. El software también puede incluir programación modular en forma de programación orientada a objeto.

Se apreciará que cualquiera de los procesos de secuenciación anteriormente descritos puede incorporarse en los métodos y/o sistemas descritos en este documento. Adicionalmente, se apreciará que pueden implementarse fácilmente otros procesos de secuenciación conocidos para su uso con los métodos y/o sistemas descritos en este documento. Se apreciará también que los métodos y sistemas descritos en este documento se diseñan para ser aplicables con cualquier tecnología de secuenciación de ácido nucleico. Adicionalmente, se apreciará que los métodos y sistemas descritos en este documento tiene incluso una aplicabilidad más amplia para cualquier campo en el que el rastreo y análisis de características en un espécimen con el paso del tiempo o desde diferentes perspectivas es importante. Por ejemplo, los métodos y sistemas descritos en este documento pueden aplicarse donde datos de imagen obtenidos por tecnologías de formación de imágenes de vigilancia, aéreas o por satélite y similares se adquieren en puntos de tiempo o perspectivas diferentes y analizan.

Ejemplos

EJEMPLO 1

IDENTIFICACIÓN DE BASES USANDO DISTRIBUCIÓN GAUSSIANA EN DATOS DE 2 CANALES

Se realiza identificación de bases en un sistema de secuenciación de 2 canales que ejecuta secuenciación de genoma de muestras humanas. Después de la generación de molde, se generan valores de intensidad para dos canales de imágenes separadas. Los valores de intensidad se normalizan de modo que los percentiles 5 y 95 se producen en 0 y 1, y cuatro distribuciones gaussianas se ajustan a los datos usando un algoritmo de maximización de la expectativa. Se calcula un centroide (valor de X,Y medio) para cada una de las cuatro distribuciones que corresponden a cada uno de los cuatro nucleótidos.

Identificación de bases para cada grupo se produce midiendo el valor de probabilidad calculado, que es la probabilidad de que el grupo pertenezca a cada una de las cuatro distribuciones. El centroide asociado con el valor máximo de probabilidad se selecciona como la identificación de bases. Este proceso de identificación de bases se realiza para cada uno de los grupos en el conjunto de datos para cada ciclo.

El término comprendiendo/que comprende se concibe en este documento para ser abierto, incluyendo no únicamente los elementos citados, sino incluyendo adicionalmente cualquier elemento adicional.

Un número de las realizaciones se han descrito. Sin embargo, se entenderá que pueden hacerse diversas modificaciones.

Las Figuras 6-9 incluyen diagramas de flujo que ilustran uno o más métodos. La Figura 6 ilustra un método 100 de

acuerdo con una realización. El método 100 puede ser, por ejemplo, un método de evaluación de la calidad de una identificación de bases a partir de una lectura de secuenciación. El método 100 puede incluir recibir, en 102, una lectura de secuenciación que tiene un número de identificaciones de bases. El método 100 también puede incluir calcular, en 104, un conjunto de valores de predictor para una identificación de bases y usar, en 106, los valores de predictor para buscar una puntuación de calidad (o métricas similares) en una tabla de calidad (o base de datos).

En un aspecto, la lectura de secuenciación utiliza identificación de bases de dos canales.

En otro aspecto, la lectura de secuenciación utiliza identificación de bases de un canal.

En otro aspecto, la tabla de calidad se genera usando puntuación Phred en un conjunto de datos de calibración. El conjunto de calibración es representativo de la variabilidad de ejecución y secuencia. En algunas realizaciones, el método 100 puede incluir generar la tabla de calidad.

En otro aspecto, los valores de predictor se seleccionan a partir del grupo que consiste en: solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; y solapamiento de señal con fondo (SOWB). En realizaciones particulares, el conjunto de valores de predictor comprende solapamiento en línea; pureza; ajuste de fase; y start5. En realizaciones particulares, el conjunto de valores de predictor comprende puntuación de hexámeros; y acumulación de motivos.

En otro aspecto, el método también incluye las etapas de descontar, en 108, puntuaciones de calidad no fiables al final de cada lectura. El método 100 también puede incluir identificar, en 110, lecturas en las que la segunda peor depuración en las primeras 25 identificaciones de bases está por debajo de un umbral preestablecido y marcar las lecturas como datos de baja calidad.

En otro aspecto, el descuento, en 108, puede incluir usar un algoritmo para identificar un umbral de fiabilidad.

En otro aspecto, identificaciones de bases fiables incluyen valores q, u otros valores indicativos de calidad de datos o significancia estadística, por encima del umbral e identificaciones de bases no fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por debajo del umbral.

En otro aspecto, el algoritmo comprende un algoritmo de Segmentos de Puntuación Máxima de Extremo Anclado (EAMSS).

En otro aspecto, el algoritmo usa un Modelo Oculto de Markov que identifica desplazamientos en las distribuciones locales de puntuaciones de calidad.

En una realización, un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación se proporciona. El sistema incluye un procesador, una capacidad de almacenamiento, y un programa para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación. El programa incluye instrucciones para (a) calcular un conjunto de valores de predictor para la identificación de bases y (b) usar los valores de predictor para buscar una puntuación de calidad en una tabla de calidad.

En otro aspecto, la lectura de secuenciación utiliza identificación de bases de dos canales.

En otro aspecto, la lectura de secuenciación utiliza identificación de bases de un canal.

En otro aspecto, la tabla de calidad se genera usando puntuación Phred en un conjunto de datos de calibración, siendo el conjunto de calibración representativo de variabilidad de ejecución y secuencia.

En otro aspecto, los valores de predictor se seleccionan a partir del grupo que consiste en: solapamiento en línea; pureza; ajuste de fase; start5; puntuación de hexámeros; acumulación de motivos; resistencia; homopolímero aproximado; decaimiento de intensidad; penúltima depuración; y solapamiento de señal con fondo (SOWB). Opcionalmente, el conjunto de valores de predictor comprende solapamiento en línea; pureza; ajuste de fase; y start5. Opcionalmente, el conjunto de valores de predictor comprende puntuación de hexámeros; y acumulación de motivos.

En otro aspecto, el programa también incluye instrucciones para (c) descontar puntuaciones de calidad no fiables al final de cada lectura y (d) identificar lecturas en las que la segunda peor depuración en las primeras 25 identificaciones de bases está por debajo de un umbral preestablecido y marcar las lecturas como datos de baja calidad.

En otro aspecto, la etapa (c) puede incluir usando un algoritmo para identificar un umbral de fiabilidad.

En otro aspecto, identificaciones de bases fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por encima del umbral e identificaciones de bases no fiables comprenden valores q, u otros valores indicativos de calidad de datos o significancia estadística, por debajo del umbral.

En otro aspecto, el algoritmo comprende un algoritmo de Segmentos de Puntuación Máxima de Extremo Anclado (EAMSS).

- 5 En otro aspecto, el algoritmo usa un Modelo Oculto de Markov que identifica desplazamientos en las distribuciones locales de puntuaciones de calidad.

La Figura 7 ilustra un método 120 de acuerdo con una realización. El método 120 puede incluir, por ejemplo, un método de generación de un valor de intensidad con corrección de ajuste de fase. El método incluye (a) realizar, en 122, una pluralidad de ciclos de una secuenciación por reacción de síntesis de tal forma que, en cada ciclo, se genera una señal indicativa de incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con lo que una porción de la señal es ruido asociado con un nucleótido incorporado durante un ciclo anterior. El método también incluye (b) detectar, en 124, la señal en cada ciclo. La señal tiene un valor de intensidad. El método 120 también incluye (c) corregir, en 126, el valor de intensidad para ajuste de fase aplicando una corrección de ajuste de fase de primer orden al valor de intensidad, en el que se calcula una nueva corrección de ajuste de fase de primer orden para cada ciclo.

En un aspecto, la corrección de ajuste de fase de primer orden comprende restar un valor de intensidad del ciclo inmediatamente anterior del valor de intensidad del ciclo actual.

- 20 En otro aspecto, el método incluye restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual.

En otro aspecto la corrección de ajuste de fase comprende:

$$25 \quad I_{(\text{ciclo})\text{corregido}} = I_{(\text{ciclo})N} - X * I_{(\text{ciclo})N-1} - Y * I_{\text{ciclo}N} + 1.$$

En otro aspecto, los valores de X y/o Y se eligen para optimizar una determinación de depuración. Opcionalmente, la determinación de depuración comprende depuración media.

- 30 En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de dos canales.

En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de un canal.

- 35 En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de cuatro canales.

En una realización, se proporciona un sistema para generar un valor de intensidad con corrección de ajuste de fase. El sistema incluye un procesador, una capacidad de almacenamiento y un programa para generar un valor de intensidad con corrección de ajuste de fase. El programa incluye instrucciones para (a) realizar una pluralidad de ciclos de una secuenciación por reacción de síntesis de tal forma que, en cada ciclo, se genera una señal indicativa de incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con lo que una porción de la señal es ruido asociado con un nucleótido incorporado durante un ciclo anterior. El programa incluye instrucciones para (b) detectar la señal en cada ciclo, en el que la señal has un valor de intensidad, y (c) corregir el valor de intensidad para ajuste de fase aplicando una corrección de ajuste de fase de primer orden al valor de intensidad. se calcula una nueva corrección de ajuste de fase de primer orden para cada ciclo.

- 45 En un aspecto, la corrección de ajuste de fase de primer orden comprende restar un valor de intensidad del ciclo inmediatamente anterior del valor de intensidad del ciclo actual.

- 50 En otro aspecto, el método incluye restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual.

En otro aspecto, la corrección de ajuste de fase comprende:

$$55 \quad I_{(\text{ciclo})\text{corregido}} = I_{(\text{ciclo})N} - X * I_{(\text{ciclo})N-1} - Y * I_{\text{ciclo}N} + 1.$$

En otro aspecto, los valores de X y/o Y se eligen para optimizar una determinación de depuración. Opcionalmente, la determinación de depuración comprende depuración media.

- 60 En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de dos canales.

En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de un canal.

En otro aspecto, la ejecución de secuenciación utiliza identificación de bases de cuatro canales.

- 65 La Figura 8 ilustra un método 140 de acuerdo con una realización. El método 140 puede ser, por ejemplo, un método de identificación de una base de nucleótido. El método 140 incluye detectar, en 142, la presencia o ausencia de una

señal en dos canales diferentes para cada una de una pluralidad de características en una matriz en un momento particular, generando de este modo un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una de las características. La combinación de valores de intensidad en cada uno de los dos canales corresponde a una de cuatro diferentes bases de nucleótidos. El método también incluye, en 144, ajustar cuatro distribuciones gaussianas a los valores de intensidad. Cada distribución tiene un centroide. El método también incluye calcular, en 146, un valor de probabilidad que indica la probabilidad de una característica particular que pertenece a cada una de las cuatro distribuciones. El método también incluye seleccionar, en 148, para cada característica de dicha pluralidad de características la distribución que tiene el mayor valor de probabilidad. Esta distribución corresponde a la identidad de la base de nucleótido presente en la característica particular.

En un aspecto, ajustar incluye usando uno o más algoritmos a partir del grupo que consiste en: un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, un algoritmo de maximización de la expectativa, y método basado en histogramas. En realizaciones particulares, ajustar incluye usando un algoritmo de maximización de la expectativa.

En otro aspecto, el método incluye normalizar los valores de intensidad.

En otro aspecto, se calcula un valor de depuración para cada característica. El valor de depuración puede ser una función de la distancia relativa desde una característica a los dos centroides más cercanos gaussianos.

En otro aspecto, se filtran características que tienen un valor de depuración por debajo de un valor umbral.

En una realización, un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación se proporciona. El sistema incluye un procesador, una capacidad de almacenamiento, y un programa para identificar una base de nucleótido. El programa incluye instrucciones para detectar la presencia o ausencia de una señal en dos canales diferentes para cada una de una pluralidad de características en una matriz en un momento particular, generando de este modo un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una de las características. La combinación de valores de intensidad en cada uno de los dos canales corresponde a una de cuatro diferentes bases de nucleótidos. El programa también incluye instrucciones para ajustar cuatro distribuciones gaussianas a los valores de intensidad. Cada distribución tiene un centroide. El programa también incluye instrucciones para calcular un valor de probabilidad que indica la probabilidad de una característica particular que pertenece a cada una de las cuatro distribuciones y seleccionar para cada característica de dicha pluralidad de características la distribución que tiene el mayor valor de probabilidad. Dicha distribución corresponde a la identidad de la base de nucleótido presente en dicha característica particular.

En un aspecto, ajustar incluye usando uno o más algoritmos a partir del grupo que consiste en: un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, un algoritmo de maximización de la expectativa, y método basado en histogramas. En realizaciones particulares, ajustar comprende usar un algoritmo de maximización de la expectativa.

En otro aspecto, el programa incluye instrucciones para normalizar los valores de intensidad.

En otro aspecto, el programa incluye instrucciones para calcular un valor de depuración para cada característica. El valor de depuración puede ser una función de la distancia relativa desde una característica a los dos centroides más cercanos gaussianos. Opcionalmente, se filtran características que tienen un valor de depuración por debajo de un valor umbral.

La Figura 9 ilustra un método 160 de acuerdo con una realización. El método 160 puede ser, por ejemplo, un método de identificación de una base de nucleótido. El método 160 incluye obtener, en 162, un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una de una pluralidad de características en una matriz. El valor de intensidad para cada característica en uno o ambos conjuntos corresponde a la presencia o ausencia de una base de nucleótido particular de entre cuatro posibles bases de nucleótidos en la característica. El método también incluye ajustar, en 164, cuatro distribuciones gaussianas a los valores de intensidad. Cada distribución tiene un centroide. El método también incluye calcular, en 166, cuatro valores de probabilidad para cada característica, en el que cada valor de probabilidad indica la probabilidad de una característica particular que pertenece a una de las cuatro distribuciones. El método también incluye seleccionar, en 168, para cada característica de dicha pluralidad de características la distribución que tiene el mayor de los cuatro valores de probabilidad. La distribución corresponde a la identidad de la base de nucleótido presente en la característica particular.

En un aspecto, ajustar incluye usando uno o más algoritmos a partir del grupo que consiste en: un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, un algoritmo de maximización de la expectativa, y método basado en histogramas. En realizaciones particulares, ajustar incluye usando un algoritmo de maximización de la expectativa.

En otro aspecto, el método también incluye normalizar los valores de intensidad.

En otro aspecto, se calcula un valor de depuración para cada característica. El valor de depuración puede ser una función de la distancia relativa desde una característica a los dos centroides más cercanos gaussianos. Opcionalmente, se filtran características que tienen un valor de depuración por debajo de un valor umbral.

5 En una realización, un sistema para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación se proporciona. El sistema incluye un procesador, una capacidad de almacenamiento, y un programa para identificar una base de nucleótido. El programa incluye instrucciones para obtener un primer conjunto de valores de intensidad y un segundo conjunto de valores de intensidad para cada una pluralidad de características en una matriz. El valor de intensidad para cada característica en uno o ambos conjuntos corresponde a la presencia o
10 ausencia de una base de nucleótido particular de entre cuatro posibles bases de nucleótidos en la característica. El programa incluye instrucciones para ajustar cuatro distribuciones gaussianas a los valores de intensidad. Cada distribución tiene un centroide. El programa incluye instrucciones para calcular cuatro valores de probabilidad para cada característica, en el que cada valor de probabilidad indica la probabilidad de una particular característica que pertenece a una de las cuatro distribuciones. El programa incluye instrucciones para seleccionar para cada
15 característica de dicha pluralidad de características la distribución que tiene el mayor de los cuatro valores de probabilidad, en el que la distribución corresponde a la identidad de la base de nucleótido presente en la característica particular.

20 En un aspecto, ajustar incluye usando uno o más algoritmos a partir del grupo que consiste en: un algoritmo de agrupamiento de k -medias, un algoritmo de agrupamiento similar a k -medias, un algoritmo de maximización de la expectativa, y método basado en histogramas. En realizaciones particulares, ajustar incluye usando un algoritmo de maximización de la expectativa.

25 En otro aspecto, el programa incluye instrucciones normalizar los valores de intensidad.

En otro aspecto, se calcula un valor de depuración para cada característica. Opcionalmente, el valor de depuración es una función de la distancia relativa desde una característica a los dos centroides más cercanos gaussianos. Opcionalmente, se filtran características que tienen un valor de depuración por debajo de un valor umbral.

30 La Figura 10 ilustra un sistema 200 formado de acuerdo con una realización que puede usarse para efectuar diversos métodos expuestos en este documento. Por ejemplo, el sistema 200 puede usarse para efectuar uno o más de los métodos 100 (Figura 6), 120 (Figura 7), 140 (Figura 8) o 160 (Figura 9). El sistema 200 pueden automatizar diversas etapas, tales como secuenciación, mientras que una o más etapas pueden realizarse manualmente o requerir de otra manera interacción de usuario. En realizaciones particulares, el usuario puede proporcionar una muestra (por ejemplo, sangre, saliva, pelo, semen, etc.) y el sistema 200 puede automáticamente preparar, secuenciar y analizar la muestra y proporcionar un perfil genético de la fuente o fuentes de la muestra. En algunas realizaciones, el sistema 200 es un sistema autónomo integrado que se ubica en un sitio. En otras realizaciones, uno o más componentes del sistema se ubican remotamente entre sí.

40 Como se muestra, el sistema 200 incluye un generador de muestra 202, un secuenciador 204 y un analizador de muestra 206. El generador de muestra 202 puede preparar la muestra para un protocolo de secuenciación designado. Por ejemplo, el generador de muestra puede preparar la muestra para SBS. El secuenciador 204 puede llevar a cabo la secuenciación para generar los datos de secuenciación. Como se ha descrito anteriormente, los datos de secuenciación pueden incluir una pluralidad de lecturas de secuenciación que incluyen numerosas identificaciones de
45 bases.

El analizador de muestra 206 puede recibir los datos de secuenciación del secuenciador 204. La Figura 10 incluye un diagrama de bloques de un analizador de muestra 206 formado de acuerdo con una realización. El analizador de muestra 206 puede usarse, por ejemplo, para analizar lecturas de secuenciación para proporcionar identificaciones de bases. El analizador de muestra 206 incluye un controlador de sistema 212 y una interfaz de usuario 214. El controlador de sistema 212 se acopla comunicativamente a la interfaz de usuario 214 y también puede acoplarse comunicativamente al secuenciador 204 y/o al generador de muestra 202.

55 En una realización ilustrativa, el controlador de sistema 212 incluye uno o más procesadores/módulos configurados para procesar y, opcionalmente, analizar datos de acuerdo con uno o más métodos expuestos en este documento. Por ejemplo, el controlador de sistema 212 puede incluir uno o más módulos configurados para ejecutar un conjunto de instrucciones que se almacenan en uno o más elementos de almacenamiento (por ejemplo, instrucciones almacenadas en un medio de almacenamiento legible por ordenador no transitorio y/o tangible, excluyendo señales) para procesar los datos de secuenciación. El conjunto de instrucciones puede incluir diversas órdenes que ordenan al controlador de sistema 212 como una máquina de procesamiento para realizar operaciones específicas tales como los flujos de trabajo, procesos y métodos descritos en este documento. A modo de ejemplo, el analizador de muestra 206 puede ser o incluir un ordenador de sobremesa, portátil, ordenador de mano, ordenador de tableta o teléfono inteligente. La interfaz de usuario 214 puede incluir hardware, firmware, software o una combinación de los mismos que habilita que un individuo (por ejemplo, un usuario) controle directa o indirectamente operación del controlador de
60 sistema 212 y los diversos componentes del mismo.
65

En la realización ilustrada, el controlador de sistema 212 incluye una pluralidad de módulos o submódulos que controlan la operación del controlador de sistema 212. Por ejemplo, el controlador de sistema 212 puede incluir los módulos 221-223 y un sistema de almacenamiento (o capacidad de almacenamiento) 226 que comunica con al menos algunos de los módulos 221-223. Los módulos 221-223 pueden ser programas en algunas realizaciones. Los módulos incluyen un módulo de corrección de ajuste de fase 221, un módulo de evaluación de calidad 222 y un módulo de identificación de base 223. El sistema 200 puede incluir otros módulos o submódulos de los módulos que se configuran para realizar las operaciones descritas en este documento. El módulo de corrección de ajuste de fase 221 se configura para generar un valor de intensidad con corrección de ajuste de fase como se expone en este documento. El módulo de evaluación de calidad 222 se configura para evaluar la calidad de una identificación de bases a partir de una lectura de secuenciación como se expone en este documento. El módulo de identificación de base 223 se configura para identificar una base de nucleótido como se expone en este documento.

Como se usa en este documento, los términos "módulo", "sistema", o "controlador de sistema" pueden incluir un sistema de hardware y/o software y circuitería que opera para realizar una o más funciones. Por ejemplo, un módulo, sistema o controlador de sistema puede incluir un procesador informático, controlador u otro dispositivo basado en lógica que realiza operaciones basándose en instrucciones almacenadas en un medio de almacenamiento legible por ordenador tangible y no transitorio, tal como una memoria de ordenador. Como alternativa, un módulo, sistema o controlador de sistema puede incluir un dispositivo por cable que realiza operaciones basándose en lógica y circuitería por cable. El módulo, sistema o controlador de sistema mostrados en las figuras adjuntas pueden representar el hardware y circuitería que opera basándose en software o instrucciones por cable, el software que dirige hardware para realizar las operaciones, o una combinación de los mismos. El módulo, sistema o controlador de sistema pueden incluir o representar circuitos o circuitería de hardware que incluyen y/o se conectan con uno o más procesadores, tal como uno o más microprocesadores informáticos.

Como se usa en este documento, los términos "software" y "firmware" son intercambiables, e incluyen cualquier programa informático almacenado en memoria para su ejecución por un ordenador, incluyendo memoria RAM, memoria ROM, memoria EPROM, memoria EEPROM y memoria RAM no volátil (NVRAM). Los tipos de memoria anteriores son únicamente ilustrativos, y por lo tanto no son limitantes como para los tipos de memoria usables para el almacenamiento de un programa informático.

En algunas realizaciones, una unidad de procesamiento, procesador, módulo o sistema informático que se "configura para" realizar una tarea u operación puede entenderse como que se estructura particularmente para realizar la tarea u operación (por ejemplo, teniendo uno o más programas o instrucciones almacenados en el mismo o usados en conjunción con el mismo personalizados o concebidos para realizar la tarea u operación, y/o teniendo una disposición de circuitería de procesamiento personalizada o concebida para realizar la tarea u operación). Para los fines de claridad y la evitación de duda, un ordenador de fin general (que puede "configurarse para" realizar la tarea u operación si se programa apropiadamente) no se "configura para" realizar una tarea u operación a no ser o hasta que se programe específicamente o modifique estructuralmente para realizar la tarea u operación.

Además, las operaciones de los métodos descritos en este documento pueden ser suficientemente complejas de tal forma que las operaciones no pueden realizarse mentalmente por un ser humano medio o un experto en la materia dentro de un periodo de tiempo comercialmente razonable. Por ejemplo, los métodos pueden basarse en cálculos relativamente complejos de tal forma que una persona de este tipo no puede completar los métodos dentro de un tiempo comercialmente razonable.

REIVINDICACIONES

1. Un método que comprende:

- 5 (a) realizar una pluralidad de ciclos de una secuenciación por reacción de síntesis, de tal forma que, en cada ciclo, se genera una señal indicativa de la incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con lo que una porción de la señal es ruido asociado a ajuste o preajuste de fase;
- 10 (b) detectar la señal en cada ciclo, en donde detectar la señal en cada ciclo incluye, detectar un valor de intensidad de la señal en un primer canal y detectar un valor de intensidad de la señal en un segundo canal; y
- 15 (c) realizar correcciones de ajuste de fase ciclo a ciclo, aplicando una nueva corrección de ajuste de fase de primer orden en cada ciclo a los valores de intensidad;
- en donde la nueva corrección de ajuste de fase de primer orden se calcula para cada ciclo, en donde la nueva corrección de ajuste de fase de primer orden para cada ciclo incluye, restar un valor de intensidad del ciclo inmediatamente anterior de un valor de intensidad del ciclo actual y también incluye, restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual; y
- en donde el nucleótido incorporado en la pluralidad de polinucleótidos idénticos, se identifica basándose en una combinación de los valores de intensidad detectados en el primer y segundo canales.

20 2. El método de la reivindicación 1, en el que se detecta un primer tipo de nucleótido en el primer canal, se detecta un segundo tipo de nucleótido en el segundo canal, se detecta un tercer tipo de nucleótido tanto en el primer como segundo canales, y no se detecta o se detecta mínimamente un cuarto tipo de nucleótido por parte del primer y segundo canales.

25 3. El método de la reivindicación 1 o de la reivindicación 2, en el que la identificación de bases de dos canales se realiza extrayendo datos de imagen de únicamente dos canales.

30 4. El método de una cualquiera de las reivindicaciones 1-3, en el que la nueva corrección de ajuste de fase de primer orden incluye pesos de ajuste y preajuste de fase y el método comprende adicionalmente elegir el peso de ajuste de fase X y el peso de preajuste de fase Y, en donde los pesos de ajuste y preajuste de fase X e Y elegidos optimizan la depuración media.

35 5. El método de una cualquiera de las reivindicaciones 1-4, en el que la identificación de bases de dos canales incluye, ajustar cuatro distribuciones gaussianas a un conjunto de datos de datos de intensidad de dos canales, de tal forma que se aplica una distribución para cada uno de los cuatro nucleótidos representados en el conjunto de datos.

40 6. El método de la reivindicación 5, en el que las distribuciones gaussianas se ajustan al conjunto de datos usando un algoritmo de grupo, en donde, más preferentemente, los polinucleótidos forman una pluralidad de grupos, teniendo cada uno de los grupos un par de valores de intensidad X, Y, haciendo referencia X e Y al primer y segundo canales, respectivamente, en donde para cada par de valores de intensidad X, Y se genera un valor de probabilidad que representa la probabilidad de que un cierto par de valores de intensidad X, Y pertenezca a una de las cuatro distribuciones, y en donde, más preferentemente, cada par de valores de intensidad X, Y tiene cuatro valores de probabilidad, uno para cada uno de los cuatro tipos de nucleótidos, indicando el máximo de los cuatro valores de probabilidad la identidad del nucleótido incorporado.

45 7. El método de la reivindicación 5, que comprende adicionalmente filtrar puntos de datos de depuración baja determinando una depuración del punto de datos como una función de distancias relativas a centroides gaussianos.

50 8. El método de una cualquiera de las reivindicaciones 1-7, en el que el valor de intensidad del ciclo inmediatamente anterior es $(X * I_{(\text{ciclo}) N-1})$ y el valor de intensidad del ciclo inmediatamente posterior es $(Y * I_{(\text{ciclo}) N+1})$, en donde X e Y representan pesos de ajuste y preajuste de fase, respectivamente.

9. El método de la reivindicación 8, en el que X e Y se eligen para optimizar una determinación de depuración.

55 10. El método de una cualquiera de las reivindicaciones 1-3, en el que la nueva corrección de ajuste de fase de primer orden comprende:

$$I_{(\text{ciclo})\text{corregido}} = I_{(\text{ciclo})N} - A * I_{(\text{ciclo})N-1} - B * I_{(\text{ciclo})N+1};$$

60 en el que las constantes A y B se calculan a partir de la estimación de tasas de ajuste y preajuste de fase y ponderan por número de ciclo; y, preferentemente, comprendiendo el método adicionalmente, optimizar sobre A y B en cada ciclo usando una búsqueda de patrones.

11. Un sistema que comprende:

- 65 un procesador;
- una capacidad de almacenamiento; y

un programa para generar un valor de intensidad con corrección de ajuste de fase, comprendiendo el programa instrucciones para:

- 5 (a) realizar una pluralidad de ciclos de una secuenciación por reacción de síntesis de tal forma que, en cada ciclo, se genera una señal indicativa de incorporación del mismo nucleótido en una pluralidad de polinucleótidos idénticos, con lo que una porción de la señal es ruido asociado a ajuste o preajuste de fase;
- (b) detectar la señal en cada ciclo, en donde detectar la señal en cada ciclo incluye, detectar un valor de intensidad de la señal en un primer canal y detectar un valor de intensidad de la señal en un segundo canal; y
- 10 (c) realizar correcciones de ajuste de fase ciclo a ciclo, aplicando una nueva corrección de ajuste de fase de primer orden en cada ciclo a los valores de intensidad;

en donde la nueva corrección de ajuste de fase de primer orden se calcula para cada ciclo, en donde la nueva corrección de ajuste de fase de primer orden para cada ciclo incluye, restar un valor de intensidad del ciclo inmediatamente anterior de un valor de intensidad del ciclo actual y también incluye, restar un valor de intensidad del ciclo inmediatamente posterior del valor de intensidad del ciclo actual; y

15 en donde el nucleótido incorporado en la pluralidad de polinucleótidos idénticos, se identifica basándose en una combinación de los valores de intensidad detectados en el primer y segundo canales.

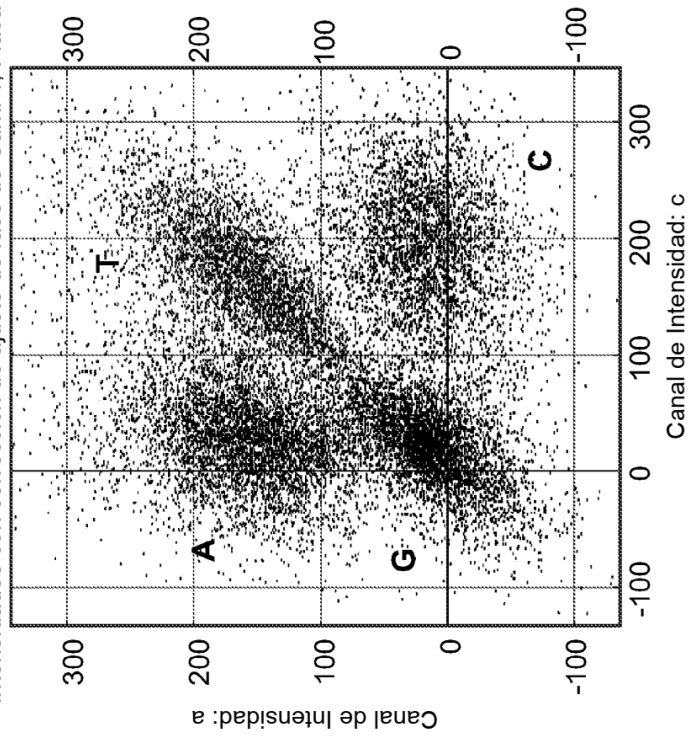
20 12. El sistema de la reivindicación 11, en el que se detecta un primer tipo de nucleótido en el primer canal, se detecta un segundo tipo de nucleótido en el segundo canal, se detecta un tercer tipo de nucleótido tanto en el primer como segundo canales, y no se detecta o se detecta mínimamente un cuarto tipo de nucleótido por parte del primer y segundo canales.

25 13. El sistema de la reivindicación 11 o de la reivindicación 12, en el que, para cada ciclo, la identificación de bases de dos canales se realiza extrayendo datos de imagen de únicamente dos canales.

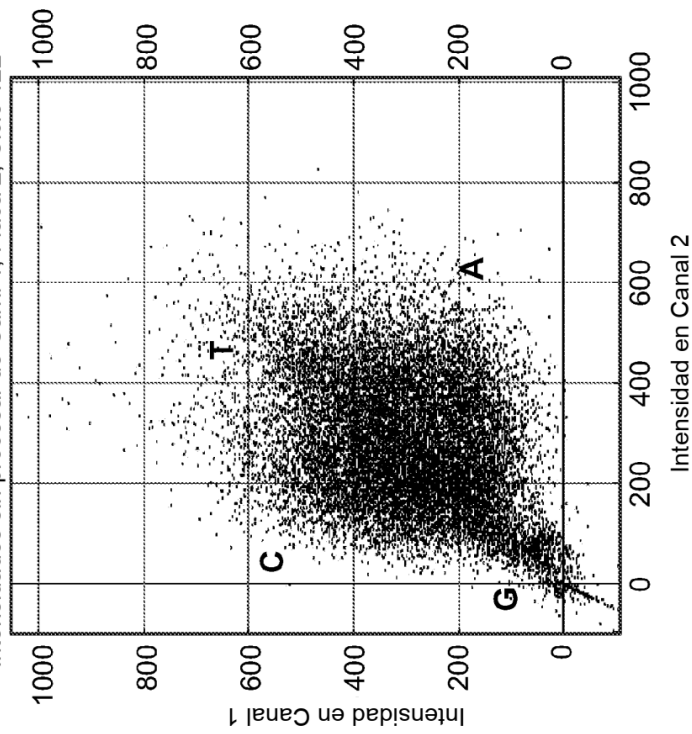
30 14. El sistema de una cualquiera de las reivindicaciones 11-13, en el que la nueva corrección de ajuste de fase de primer orden incluye pesos de ajuste y preajuste de fase e incluyendo el programa instrucciones para elegir el peso de ajuste de fase X y el peso de preajuste de fase Y, en donde los pesos de ajuste y preajuste de fase X e Y elegidos optimizan la depuración media.

15. El sistema de la reivindicación 14, en el que los pesos de ajuste y preajuste de fase se seleccionan usando una búsqueda de patrones sobre los posibles pesos de ajuste y preajuste de fase.

Intensidades con corrección de ajuste de fase de Carril 1, Placa 2, Ciclo 122



Intensidades sin procesar de Carril 1, Placa 2, Ciclo 122



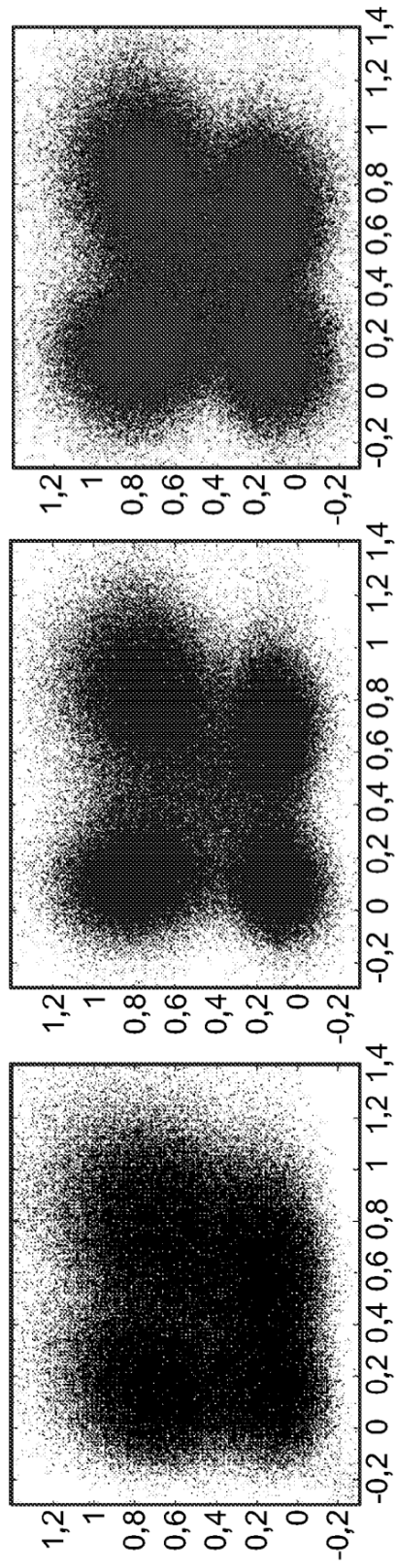


FIG. 2A

FIG. 2B

FIG. 2C

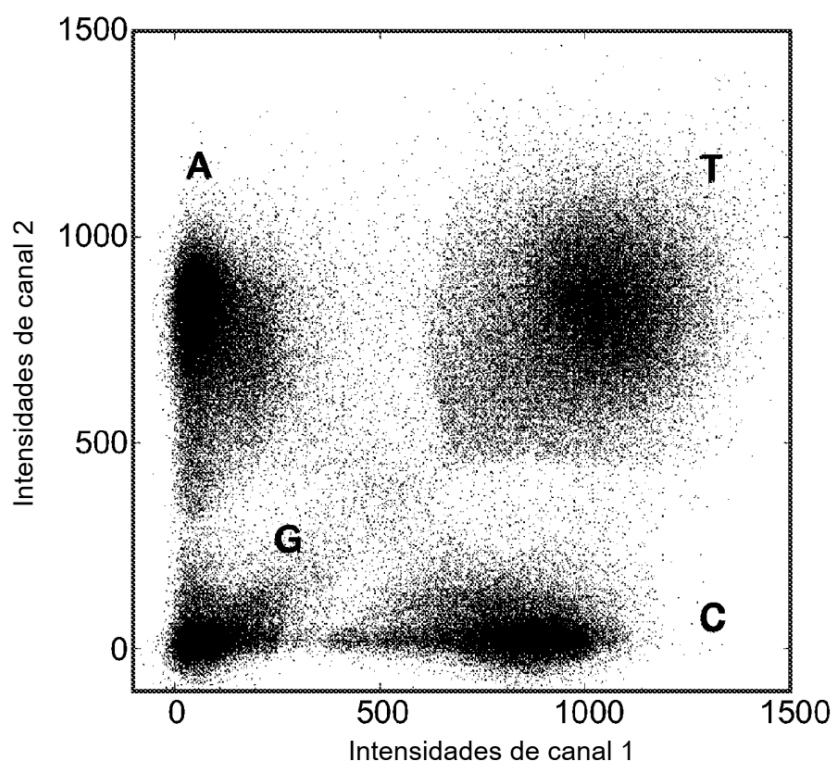


FIG. 3

Identificación de base de 2 canales

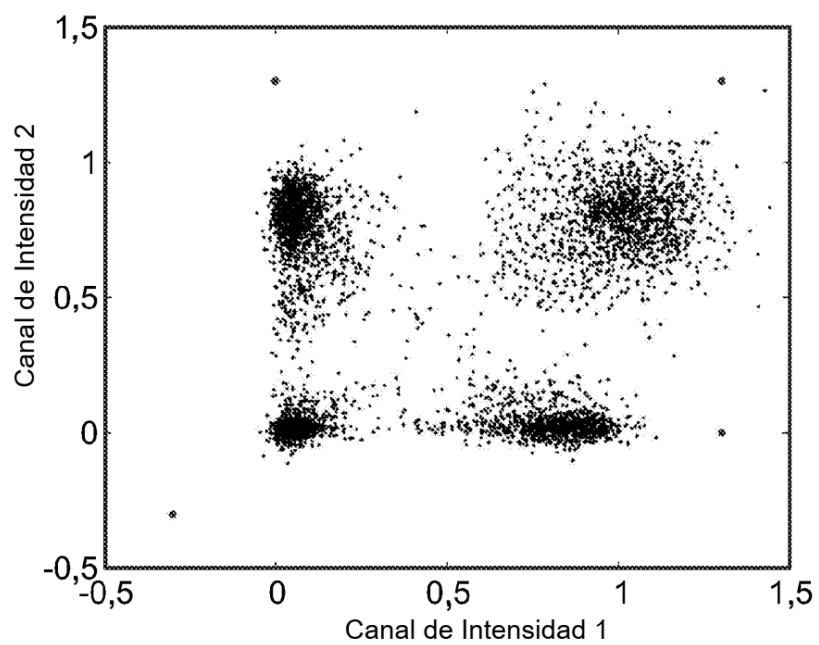
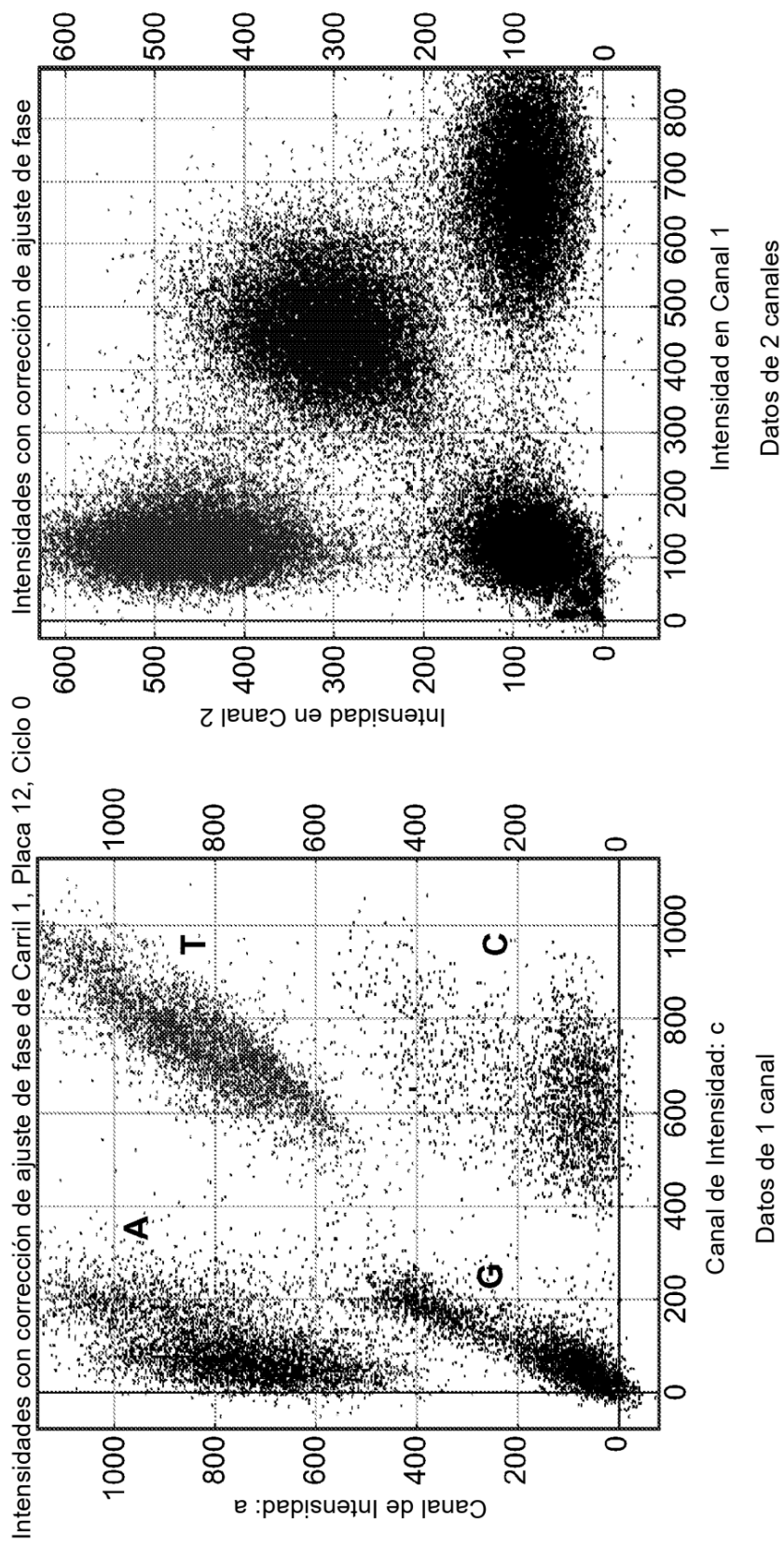


FIG. 4



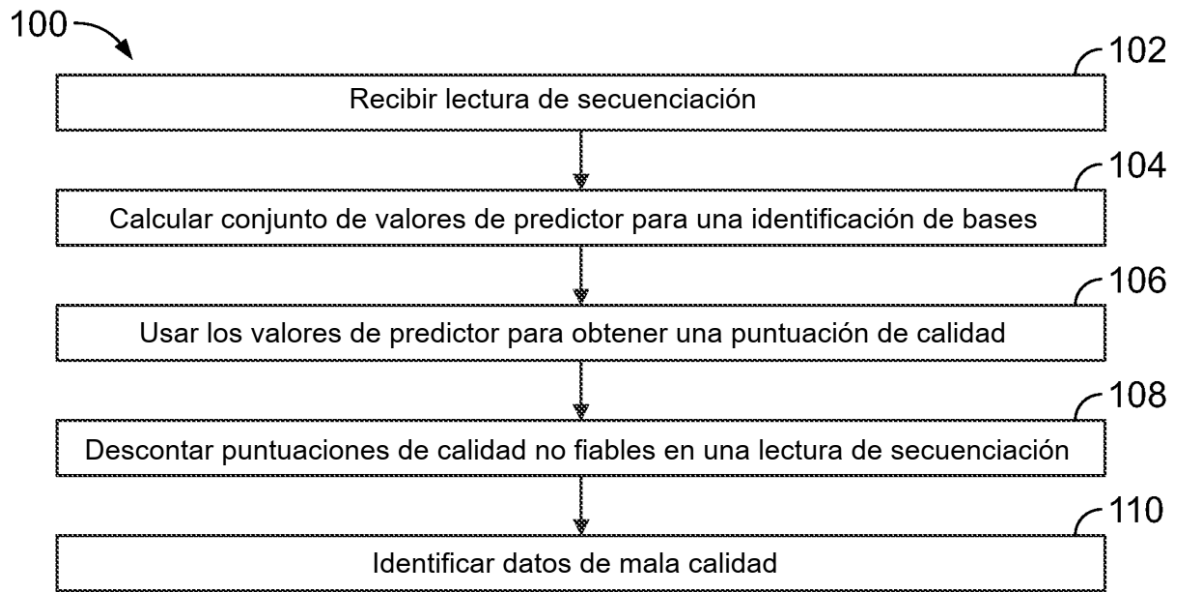


FIG. 6

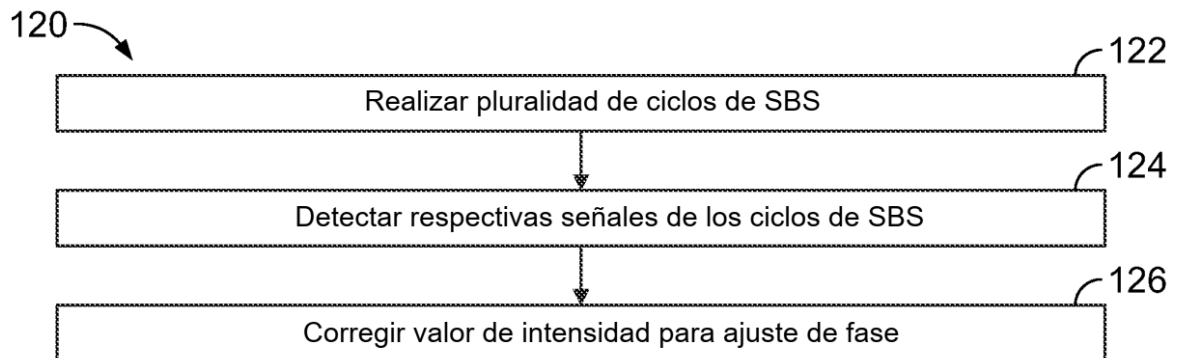


FIG. 7

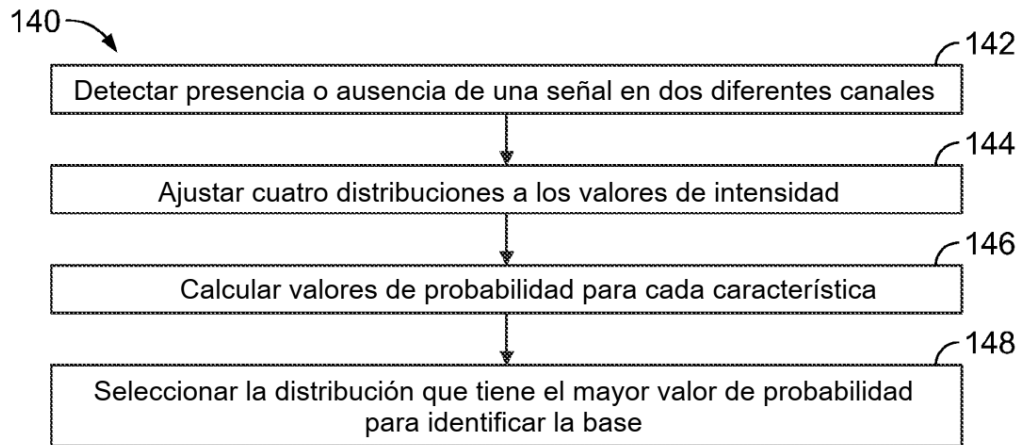


FIG. 8

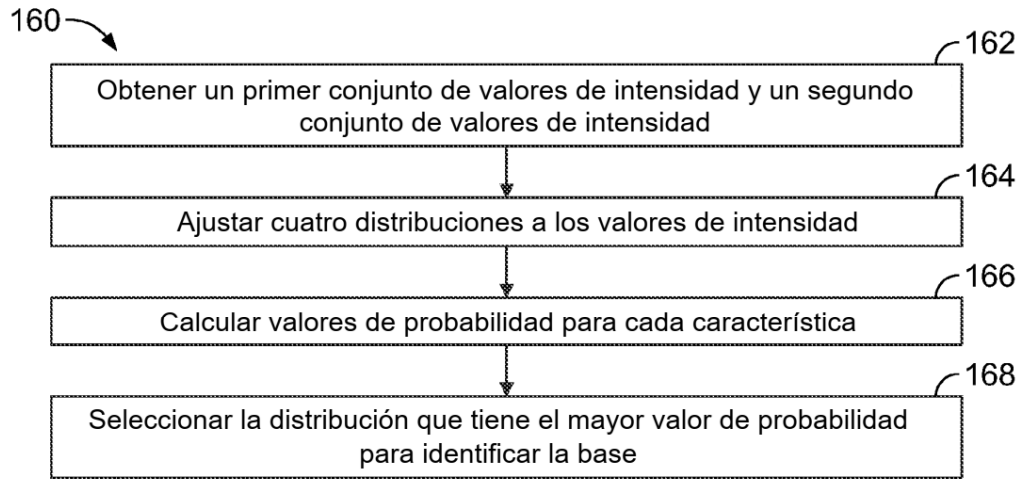


FIG. 9

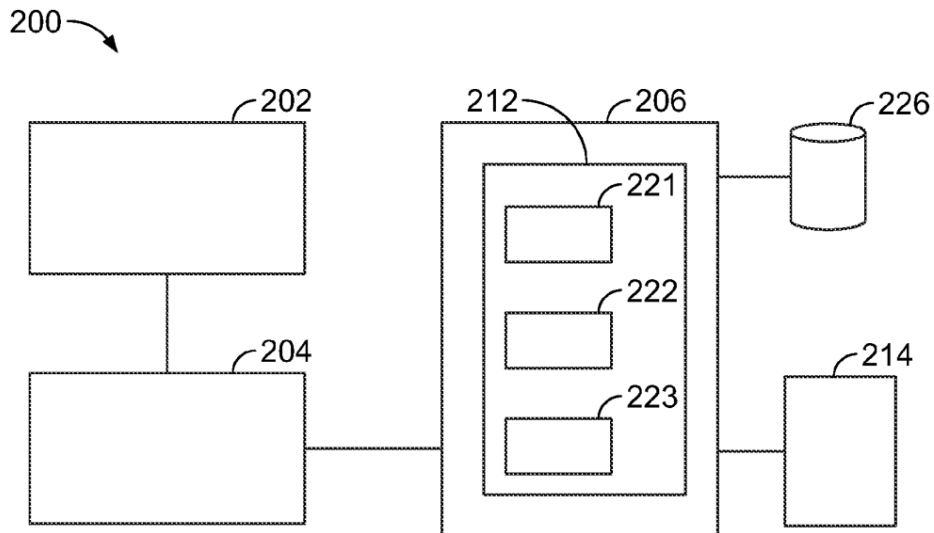


FIG. 10