

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 917 403**

51 Int. Cl.:

C12Q 1/6869 (2008.01)

G16B 25/00 (2009.01)

G16B 30/00 (2009.01)

G16B 30/10 (2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **05.01.2018** **PCT/US2018/012580**

87 Fecha y número de publicación internacional: **12.07.2018** **WO18129314**

96 Fecha de presentación y número de la solicitud europea: **05.01.2018** **E 18702831 (1)**

97 Fecha y número de publicación de la concesión europea: **20.04.2022** **EP 3566158**

54 Título: **Corrección de ajuste de fase**

30 Prioridad:

06.01.2017 US 201762443294 P

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

08.07.2022

73 Titular/es:

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

72 Inventor/es:

LANGLOIS, ROBERT y
BELITZ, PAUL

74 Agente/Representante:

DEL VALLE VALIENTE, Sonia

ES 2 917 403 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

Corrección de ajuste de fase

5 **Antecedentes**

La descripción se refiere a secuenciación de ácidos nucleicos. Más específicamente, la descripción se refiere a sistemas y métodos para secuenciación en tiempo real con correcciones de ajuste de fase.

10 En un sitio particular en una celda de flujo u otro sustrato, se analizan juntas múltiples copias de una molécula de ácido nucleico, que tienen todas la misma secuencia (posiblemente con variaciones limitadas introducidas involuntariamente por el procesamiento de la muestra). Se usan copias suficientes para garantizar que se produzca una señal suficiente para permitir una llamada de base fiable. La colección de moléculas de ácido nucleico en un sitio se denomina agrupación.

15 El ajuste de fase representa un artefacto no deseado que surge de la secuenciación de múltiples moléculas de ácido nucleico dentro de una agrupación. El ajuste de fase es la velocidad a la que las señales, tales como la fluorescencia de moléculas individuales dentro de una agrupación, pierden sincronización entre sí. A menudo, el término ajuste de fase se reserva para la señal de contaminación procedente de algunas moléculas que se encuentran detrás, y el término preajuste de fase se usa para la señal de contaminación procedente de otras moléculas que van por delante. El ajuste de fase y el preajuste de fase juntos describen el rendimiento del aparato de secuenciación y la química. El documento WO 2015/084985 describe métodos y sistemas para el análisis de datos de imagen generados a partir de diversos puntos de referencia, que son útiles para el análisis en tiempo real de datos de imagen y secuencia generados durante metodologías de secuenciación de ADN.

25 **Resumen**

Determinados aspectos de esta descripción se refieren a métodos para determinar valores de color corregidos a partir de datos de imagen adquiridos por un secuenciador de ácidos nucleicos durante un ciclo de llamada de base, donde el secuenciador incluye un sistema de adquisición de imágenes, uno o más procesadores y una memoria. Tales métodos pueden caracterizarse por las siguientes operaciones: (a) obtener una imagen de un sustrato (por ejemplo, una parte de una celda de flujo) que incluye una pluralidad de sitios donde se leen las bases de ácidos nucleicos; (b) medir valores de color de la pluralidad de sitios a partir de la imagen del sustrato; (c) almacenar los valores de color en una memoria intermedia de procesador del uno o más procesadores; (d) recuperar valores de color con corrección de fase parcial de la pluralidad de sitios, donde los valores de color con corrección de fase parcial se almacenaron en la memoria del secuenciador durante un ciclo de llamada de base inmediatamente anterior; (e) determinar una corrección de preajuste de fase; y (f) determinar los valores de color corregidos. En diversas implementaciones, estas operaciones se realizan todas durante un único ciclo de llamada de base. En determinadas realizaciones, los métodos incluyen adicionalmente usar los valores de color corregidos para realizar llamadas de base para la pluralidad de sitios.

45 Durante la secuenciación, los sitios muestran colores que representan tipos de bases de ácidos nucleicos. Los valores de color medidos y almacenados pueden ser valores de intensidad u otra magnitud en una longitud de onda o intervalo de longitudes de onda particular. En algunas implementaciones, los valores de color se determinan a partir de solo dos canales del secuenciador. En algunas implementaciones, los valores de color se obtienen a partir de cuatro canales del secuenciador. Aunque esta descripción se centra en la corrección de ajuste de fase de señales de color, los conceptos se aplican a otros tipos de señales generadas durante la secuenciación de agrupaciones de ácidos nucleicos que tienen secuencias idénticas. Los ejemplos de tales otras señales incluyen radiación fuera del espectro visible, concentración de iones, etc.

50 En determinadas realizaciones, determinar los valores de color corregidos en (f) usa (i) los valores de color en la memoria intermedia de procesador, (ii) los valores con corrección de fase parcial almacenados durante el ciclo inmediatamente anterior, y (iii) la corrección de preajuste de fase. En determinadas realizaciones, determinar la corrección de preajuste de fase en (e) usa (i) los valores de color con corrección de fase parcial almacenados durante el ciclo de llamada de base inmediatamente anterior, y (ii) los valores de color almacenados en la memoria intermedia de procesador.

55 En determinadas realizaciones, la corrección de preajuste de fase incluye una ponderación. En tales realizaciones, la operación de determinar los valores de color corregidos puede incluir multiplicar la ponderación por los valores de color de la pluralidad de sitios medidos a partir de la imagen del sustrato.

60 En determinadas implementaciones, los métodos incluyen adicionalmente determinar una corrección de ajuste de fase para el ciclo de llamada de base inmediatamente posterior. Como ejemplo, determinar la corrección de ajuste de fase para el ciclo de llamada de base inmediatamente posterior incluye analizar (i) los valores de color con corrección de fase parcial almacenados en la memoria del secuenciador, y (ii) los valores de color almacenados en la memoria intermedia de procesador. En determinadas realizaciones que incluyen determinar una corrección de ajuste de fase para el ciclo de llamada de base inmediatamente posterior, los métodos incluyen adicionalmente (i)

producir valores de color con corrección de fase parcial para el ciclo de llamada de base inmediatamente posterior aplicando la corrección de ajuste de fase a valores de color de la pluralidad de sitios almacenados en la memoria de secuenciador; y (ii) almacenar los valores de color con corrección de fase parcial para el ciclo de llamada de base inmediatamente posterior en la memoria del secuenciador. En determinadas realizaciones, la producción de los valores de color con corrección de fase parcial para el ciclo de llamada de base inmediatamente posterior incluye adicionalmente sumar (i) los valores de color con corrección de ajuste de fase de la pluralidad de sitios, y (ii) los valores de color de la pluralidad de sitios de la imagen del sustrato medido en (b). En algunas implementaciones, almacenar los valores de color con corrección de fase parcial para el ciclo de llamada de base inmediatamente posterior almacena los valores de color parcialmente corregidos en memorias intermedias en tesela de la memoria del secuenciador.

En determinadas realizaciones, los métodos se realizan en tiempo real durante la adquisición de lecturas de secuencia por el secuenciador de ácidos nucleicos. En determinadas realizaciones, el secuenciador de ácidos nucleicos secuencia mediante la síntesis de ácidos nucleicos en la pluralidad de sitios. En determinadas realizaciones donde el sustrato incluye una celda de flujo, la celda de flujo se divide lógicamente en teselas, y cada tesela representa una región de la celda de flujo que comprende un subconjunto de sitios, subconjunto que se captura en una única imagen del sistema de adquisición de imágenes.

En algunas realizaciones que emplean tales sistemas, en la operación (d) (recuperar valores de color con corrección de fase parcial de la pluralidad de sitios), los valores de color con corrección de fase parcial se almacenaron previamente en memorias intermedias en tesela de la memoria del secuenciador, donde las memorias intermedias en tesela se designan para almacenar datos que representan imágenes en tesela individuales sobre el sustrato. En determinadas realizaciones, la memoria tiene una capacidad de almacenamiento de aproximadamente 512 Gigabytes o menos, o aproximadamente 256 Gigabytes o menos. En determinadas realizaciones, por ejemplo, la memoria tiene una capacidad de almacenamiento menor que dos veces la capacidad requerida para almacenar los datos contenidos en el número total de teselas en dos celdas de flujo. En algunas realizaciones, el procesamiento descrito en la presente memoria ahorra al menos aproximadamente 50 Gigabytes; en algunas realizaciones ahorra al menos aproximadamente 100 Gigabytes.

En algunas implementaciones, antes de la operación (a) (obtener una imagen de un sustrato), los métodos incluyen adicionalmente proporcionar reactivos a la celda de flujo y permitir que los reactivos interactúen con sitios para mostrar los colores que representan tipos de bases de ácidos nucleicos durante el ciclo de llamada de base. En tales implementaciones, el método puede incluir adicionalmente, después de la operación (f) (determinar los valores de color corregidos): (i) proporcionar reactivos nuevos a la celda de flujo y permitir que los reactivos nuevos interactúen con los sitios para mostrar colores que representan tipos de bases de ácidos nucleicos para un siguiente ciclo de llamada de base; y (ii) repetir las operaciones (a)-(e) para el siguiente ciclo de llamada de base. Tales métodos pueden incluir adicionalmente crear un primer subproceso de procesador para realizar las operaciones (a)-(f) para el ciclo de llamada de base, y crear un segundo subproceso de procesador para realizar las operaciones (a)-(f) para el siguiente ciclo de llamada de base. En determinadas realizaciones, los métodos incluyen adicionalmente asignar la memoria intermedia de procesador y una segunda memoria intermedia de procesador, donde la segunda memoria intermedia de procesador se usa para determinar los valores de color corregidos en (f).

Otros aspectos determinados de la descripción se refieren a secuenciadores de ácidos nucleicos que pueden caracterizarse por los siguientes elementos: un sistema de adquisición de imágenes; una memoria y uno o más procesadores diseñados o configurados para: (a) obtener datos que representan una imagen de un sustrato que incluye una pluralidad de sitios donde se leen las bases de ácidos nucleicos (los sitios muestran, por ejemplo, colores que representan tipos de bases de ácidos nucleicos); (b) obtener valores de color de la pluralidad de sitios a partir de la imagen del sustrato; (c) almacenar los valores de color en una memoria intermedia de procesador; (d) recuperar valores de color con corrección de fase parcial de la pluralidad de sitios para un ciclo de llamada de base (los valores de color con corrección de fase parcial se almacenaron en la memoria del secuenciador durante un ciclo de llamada de base inmediatamente anterior); (e) determinar una corrección de preajuste de fase; y (f) determinar valores de color corregidos a partir de, por ejemplo, (i) los valores de color en la memoria intermedia de procesador, (ii) los valores con corrección de fase parcial almacenados durante el ciclo inmediatamente anterior, y (iii) la corrección de preajuste de fase.

Las instrucciones u otra configuración para determinar una corrección de preajuste de fase puede incluir la configuración para determinar la corrección de preajuste de fase de (i) los valores de color con corrección de fase parcial almacenados durante el ciclo de llamada de base inmediatamente anterior, y (ii) los valores de color almacenados en la memoria intermedia de procesador.

En determinadas realizaciones, la memoria se divide en una pluralidad de memorias intermedias en tesela, cada una designada para almacenar datos que representan una única imagen de una tesela en el sustrato. En determinadas realizaciones, la memoria tiene una capacidad de almacenamiento menor de aproximadamente 550 Gigabytes (en algunos ejemplos, esto es menor que dos veces la capacidad requerida para almacenar los datos contenidos en el número total de teselas en dos celdas de flujo).

Los procesadores pueden configurarse para realizar las operaciones mencionadas de varias maneras, tales como recibir instrucciones ejecutables legibles por máquina. En algunos casos, los procesadores se programan con firmware o núcleos de procesamiento personalizados, tales como núcleos de procesamiento de señales digitales. En diversas realizaciones, el procesador o los procesadores están diseñados o configurados para realizar (y/o controlar) cualquiera o más de las operaciones del método descritas anteriormente.

En algunas implementaciones, las características de corrección de ajuste de fase descritas en la presente memoria reducen sustancialmente el coste de un instrumento de secuenciación al utilizar la memoria de manera más eficiente (por ejemplo, memoria de acceso aleatorio (RAM) . Algunas realizaciones emplean estas características de corrección de ajuste de fase en el contexto del análisis en tiempo real (RTA) en plataformas de secuenciación

Estas y otras características de la descripción se presentarán con mayor detalle a continuación, con referencia a los dibujos asociados.

Breve descripción de los dibujos

La figura 1 es un diagrama de bloques de un secuenciador con hardware para análisis en tiempo real de datos de imagen tomados de agrupaciones de ácidos nucleicos.

La figura 2 es una ilustración de datos de secuenciación de dos canales utilizados para ilustrar los conceptos de ajuste de fase y preajuste de fase.

La figura 3 representa una arquitectura de celda de flujo que incluye una pluralidad de teselas, cada una de las cuales contiene muchas agrupaciones.

La figura 4 representa una matriz de datos que contiene datos de magnitud para agrupaciones en una tesela u otra parte tomada como imagen de una celda de flujo; los datos de magnitud pueden ser valores de intensidad de luz para cada uno de dos o más canales de color.

La figura 5 representa esquemáticamente una primera configuración y metodología de procesamiento para realizar corrección de ajuste de fase en tiempo real.

La figura 6 presenta un diagrama de flujo de un procedimiento de llamada de base que puede emplear la configuración de procesador y memoria representada en la figura 5.

La figura 7 representa esquemáticamente una segunda configuración y metodología de procesamiento para realizar la corrección de ajuste de fase en tiempo real. Esta configuración reduce los requisitos de memoria de sistema.

La figura 8 representa esquemáticamente una tercera configuración y metodología de procesamiento para realizar la corrección de ajuste de fase en tiempo real. Esta configuración reduce aún más los requisitos de memoria de sistema.

La figura 9 presenta un diagrama de flujo de alto nivel de los primeros ciclos de procesamiento que pueden emplearse con la configuración de procesador y memoria de la figura 8 y, en algunas implementaciones, la figura 7.

La figura 10 presenta un diagrama de flujo de ciclos de procesamiento que conducen a la llamada de base con corrección de ajuste de fase completa. Un ciclo de este tipo puede realizarse en los ciclos de procesamiento tercero y posteriores cuando se secuencian agrupaciones de una tesela.

La figura 11 presenta datos comparativos para métodos de corrección de ajuste de fase, usando un algoritmo de memoria principal reducido.

Descripción detallada

Definiciones

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

Los títulos proporcionados en la presente memoria no pretenden limitar la descripción.

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

Los términos definidos inmediatamente a continuación se describen más completamente con referencia a la memoria descriptiva en su conjunto. Debe entenderse que esta descripción no se limita a la metodología, protocolos y reactivos particulares descritos, ya que estos pueden variar, dependiendo del contexto en que los expertos en la técnica los usen.

Como se utiliza en la presente memoria, los términos singulares “un(o)”, “una” y “el/la” incluyen la referencia plural salvo que el contexto indique claramente algo distinto. El término “pluralidad” se refiere a más de un elemento. Por ejemplo, el término se usa en la presente memoria en referencia a una serie de lecturas para producir una isla en fase usando los métodos descritos en la presente memoria.

El término “parte” se usa en la presente memoria en referencia a la cantidad de información de secuencia del genoma, cromosoma o haplotipo en una muestra biológica que, en suma, es menor que la información de secuencia de un genoma completo, un cromosoma completo o un haplotipo completo, tal como es evidente a partir del contexto.

El término “muestra” se refiere en la presente memoria a una muestra, normalmente derivada de un fluido biológico, célula, tejido, órgano u organismo, que contiene un ácido nucleico o una mezcla de ácidos nucleicos que contiene al menos una secuencia de ácido nucleico que va a secuenciarse. Tales muestras incluyen, pero no se limitan a esputo/fluido oral, líquido amniótico, líquido cefalorraquídeo, sangre, una fracción de sangre (por ejemplo, suero o plasma), muestras de biopsia con aguja fina (por ejemplo, biopsia quirúrgica, biopsia con aguja fina, etc.), orina, saliva, semen, sudor, lágrimas, líquido peritoneal, líquido pleural, líquido de lavado, explante tisular, cultivo de órganos y cualquier otro tejido o preparación celular, o fracción o derivado de los mismos o aislado de los mismos.

Aunque la muestra a menudo se toma de un sujeto humano (por ejemplo, un paciente), la muestras pueden tomarse de cualquier organismo que tenga cromosomas incluidos, aunque no de forma limitativa, perros, gatos, caballos, cabras, ovejas, ganado vacuno, cerdos, etc. La muestra puede usarse directamente tal como se obtiene de la fuente biológica o después de un pretratamiento para modificar el carácter de la muestra. Por ejemplo, dicho pretratamiento puede incluir preparar plasma a partir de sangre, diluir fluidos viscosos, etc. Los métodos de pretratamiento también pueden implicar, aunque no de forma limitativa, filtración, precipitación, dilución, destilación, mezcla, centrifugación, congelación, liofilización, concentración, amplificación, fragmentación de ácido nucleico, inactivación de componentes interferentes, la adición de reactivos, lisis, etc. Si dichos métodos de pretratamiento se emplean con respecto a la muestra, dichos métodos de pretratamiento son de forma típica tales que el(los) ácido(s) nucleico(s) de interés permanecen en la muestra de prueba, a veces en una concentración proporcional a la de una muestra de prueba no tratada (por ejemplo, a saber, una muestra que no se somete a ninguno de dicho(s) método(s) de pretratamiento). Dichas muestras “tratadas” o “procesadas” siguen siendo consideradas muestras “de prueba” biológicas con respecto a los métodos descritos en la presente memoria.

Los términos “polinucleótido”, “ácido nucleico” y “moléculas de ácido nucleico” se usan indistintamente y se refieren a una secuencia de nucleótidos unida covalentemente (es decir, ribonucleótidos para ARN y desoxirribonucleótidos para ADN) en la que la posición 3' de la pentosa de un nucleótido está unida por un grupo fosfodiéster a la posición 5' de la pentosa del siguiente. Los nucleótidos incluyen secuencias de cualquier forma de ácido nucleico incluidas, aunque no de forma limitativa, moléculas de ARN y ADN. El término “polinucleótido” incluye, sin limitación, polinucleótidos monocatenarios y bicatenarios.

Las moléculas de polinucleótidos monocatenarios pueden haberse originado en forma monocatenaria, como ADN o ARN, o haberse originado en forma de ADN bicatenario (ADNbc) (por ejemplo, segmentos de ADN genómico, productos de PCR y amplificación y similares). Por lo tanto, un polinucleótido monocatenario puede ser la cadena sentido o antisentido de un dúplex de polinucleótido. Se conocen bien en la técnica los métodos de preparación de moléculas de polinucleótidos monocatenarios adecuadas para su uso en los métodos descritos usando técnicas convencionales. La secuencia precisa de las moléculas de polinucleótidos primarios generalmente no es esencial para las realizaciones descritas y puede ser conocida o desconocida. Las moléculas de polinucleótidos monocatenarios pueden representar moléculas de ADN genómico (por ejemplo, ADN genómico humano) que incluyen secuencias tanto de intrones como de exones (secuencia codificante), así como secuencias reguladoras no codificantes tales como secuencias promotoras y potenciadoras.

El ácido nucleico descrito en la presente memoria puede ser de cualquier longitud adecuada para su uso en los métodos proporcionados. Por ejemplo, los ácidos nucleicos diana pueden tener al menos 10, al menos 20, al menos 30, al menos 40, al menos 50, al menos 75, al menos 100, al menos 150, al menos 200, al menos 250, al menos 500, o al menos 1000 kb de longitud o más.

En el contexto de una celda de flujo u otro sustrato para la secuenciación, el término “sitio” se refiere a una región pequeña donde tiene lugar la secuenciación. En muchas realizaciones, un sitio contiene múltiples, normalmente numerosas copias de una única secuencia de ácido nucleico a partir de la cual se obtienen datos de secuenciación. Los datos de secuencia obtenidos de un sitio pueden ser una “lectura”

El término “polimorfismo” o “polimorfismo genético” se usa en la presente memoria en referencia a la aparición en la misma población de dos o más alelos en un locus genético. Varias formas de polimorfismo incluyen polimorfismos de un solo nucleótido, repeticiones en tándem, microdeleciones, inserciones, indeles y otros polimorfismos.

Una “llamada de base” es una base asignada (tipo de nucleótido) a datos de secuencia para una ubicación particular en una secuencia de polinucleótidos. Un secuenciador puede emitir una llamada de base para cada posición en el ácido nucleico que se secuencia. A veces se atribuye una calidad de la llamada a una llamada de base.

El término “lectura” se refiere a una secuencia leída de una parte de una muestra de ácido nucleico. De forma típica, aunque no necesariamente, una lectura representa una secuencia corta de pares de bases contiguos en la muestra. La lectura puede representarse simbólicamente mediante la secuencia de pares de bases (en ATCG) de la parte de muestra. Puede almacenarse en un dispositivo de memoria y procesarse según sea apropiado para determinar si coincide con una secuencia de referencia o cumple otros criterios. Se puede obtener una lectura directamente de un aparato de secuenciación o indirectamente a partir de información de secuencia almacenada con respecto a la muestra. En algunos casos, una lectura es una secuencia de ADN de longitud suficiente (por ejemplo, al menos aproximadamente 25 pb) que puede usarse para identificar una secuencia o región más grande, por ejemplo, que puede alinearse y asignarse específicamente a un cromosoma o región genómica o gen.

El término “secuenciación de próxima generación (NGS)” en la presente memoria se refiere a métodos de secuenciación que permiten la secuenciación masivamente en paralelo de moléculas clonalmente amplificadas y de moléculas de ácido nucleico individuales. Los ejemplos no limitativos de NGS incluyen secuenciación por síntesis usando terminadores de colorante reversibles y secuenciación por ligación.

El término “parámetro” se refiere en la presente memoria a un valor numérico que caracteriza una propiedad física o una representación de esa propiedad. En algunas situaciones, un parámetro caracteriza numéricamente un conjunto de datos cuantitativos y/o una relación numérica entre conjuntos de datos cuantitativos. Por ejemplo, la media y la varianza de un ajuste de distribución estándar a un histograma son parámetros.

El término “umbral” se refiere en la presente memoria a cualquier número que se usa como punto de corte para caracterizar una muestra, un ácido nucleico o una parte del mismo (por ejemplo, una lectura). El umbral puede compararse con un valor medido o calculado para determinar si la fuente que da lugar a tales sugerencias de valor debe clasificarse de manera particular. Los valores de umbral pueden identificarse empírica o analíticamente. La elección de un umbral depende del nivel de confianza con que el usuario desea realizar la clasificación. A veces se eligen para un propósito particular (por ejemplo, para equilibrar la sensibilidad y la selectividad).

El análisis en tiempo real se refiere a un procedimiento y sistema en el que el procesamiento y el análisis de datos se realizan en segundo plano de la adquisición de datos durante una ejecución de secuenciación de ADN. Un ejemplo de un sistema de análisis en tiempo real se describe en la patente estadounidense n.º 8.954.076.

Contexto para el ajuste de fase

Aparato de secuenciación

La figura 1 muestra un diagrama de bloques de algunas características de un secuenciador 100 de ácidos nucleicos típico o un sistema que incluye dicho secuenciador. En particular, el sistema 100 incluye una celda 101 de flujo y un sistema 103 de adquisición de imágenes, uno o más procesadores 105 con una o más memorias 107 intermedias, y una memoria 109 de sistema (a veces denominada memoria principal) que incluye una pluralidad de memorias 111 intermedias en tesela. Normalmente, la memoria 109 de sistema se proporciona en un dispositivo que no forma parte de un circuito integrado que contiene cualquiera del uno o más procesadores 105. En determinadas realizaciones, la memoria de sistema es una memoria volátil tal como una memoria de acceso aleatorio o RAM, por ejemplo, DRAM, un disco duro de estado sólido o una unidad de disco duro.

La célula de flujo y el sistema de adquisición de imágenes contiene componentes diseñados o configurados según los principios entendidos en el campo de la secuenciación de ácidos nucleicos, y no se describirán en detalle en la presente memoria. Los sistemas de análisis de imágenes adecuados y las celdas de flujo asociadas se emplean en secuenciadores de ácidos nucleicos tales como las series de secuenciadores MiSeq y HiSeq disponibles de Illumina, Inc. de San Diego, California. Para obtener información adicional, véase la patente estadounidense n.º 8.241.573, la patente estadounidense n.º 9.193.996 y la patente estadounidense n.º 8.951.781.

En general, las secuencias de ácido nucleico adecuadas para su uso con los métodos descritos proporcionan una detección rápida y eficaz de una pluralidad de ácidos nucleico diana en paralelo. Pueden incluir componentes fluidicos capaces de administrar reactivos de amplificación y/o reactivos de secuenciación a uno o más fragmentos de ADN inmovilizados, incluyendo el sistema componentes tales como bombas, válvulas, depósitos, líneas fluidicas y similares. Una celda de flujo puede configurarse y/o usarse en un sistema integrado para la detección de ácidos nucleicos diana. Por ejemplo, se describen celdas de flujo ilustrativas en los documentos US 2010/0111768 A1 y US con n.º de serie 13/273.666.

Tal como se ejemplifica para las celdas de flujo, puede usarse uno o más de los componentes fluidicos de un sistema integrado tanto para un método de amplificación como para un método de detección. Por ejemplo, puede usarse uno o más de los componentes fluidicos de un sistema integrado para un método de amplificación y para el suministro de reactivos de secuenciación en un método de secuenciación. Como alternativa, un sistema integrado puede incluir sistemas fluidicos separados para llevar a cabo métodos de amplificación y para llevar a cabo métodos de detección.

Para los fines de esta descripción, basta con entender que la celda de flujo primero recibe e inmoviliza o captura de otro modo una muestra de ácido nucleico que va a secuenciarse y luego se expone a diversos reactivos asociados con el procedimiento de secuenciación. En determinadas realizaciones, el procedimiento de secuenciación es un procedimiento mediante secuencia por síntesis, aunque pueden emplearse otras tecnologías de secuenciación.

El sistema 103 de adquisición de imágenes incluye componentes ópticos tales como componentes de excitación de fluorescencia (por ejemplo, un láser y espejos y lentes asociados) para iluminar sitios en la celda de flujo donde tiene lugar la secuenciación y componentes de captura de imágenes para capturar imágenes de fluorescencia en partes de la celda de flujo que tienen múltiples sitios. Los datos capturados por el sistema de adquisición de imágenes contienen información adecuada para determinar qué nucleótido está leyéndose en cualquier sitio dado en cualquier ciclo de secuenciación dado.

Para permitir el análisis en tiempo real, el secuenciador 100 normalmente incluye procesadores incorporados y una memoria que interpretan y almacenan datos de imagen del sistema 103 de adquisición de imágenes. Los ejemplos de procesadores adecuados para el secuenciador incluyen la clase Xeon E5 de Intel. Normalmente, el procesador 105 incluye múltiples memorias 107 intermedias que almacenan temporalmente los datos de imagen tomados durante un único ciclo de adquisición de imágenes. En la realización representada, las memorias intermedias de procesador se asignan en la memoria de sistema. Una memoria intermedia de procesador dada puede asociarse con un subproceso de procesador particular creado para analizar datos de imagen de una región de la celda de flujo durante el análisis en tiempo real. En determinadas realizaciones, los datos de imagen analizados por un subproceso son el de una única tesela (descrita a continuación), capturados durante un único ciclo de adquisición de imágenes. En determinadas realizaciones, la memoria intermedia puede almacenar aproximadamente 400 Gigabytes de datos. Tal como se usa en la presente memoria, un subproceso es una secuencia ordenada de instrucciones que le dice al procesador qué operaciones ejecutar. Las instrucciones configuran el procesador utilizando código de máquina ejecutable seleccionado de un conjunto de instrucciones de lenguaje de máquina específico, o “instrucciones nativas” diseñadas en el procesador de hardware.

El conjunto de instrucciones del lenguaje de máquina, o el conjunto de instrucciones nativas, lo conoce, y está esencialmente integrado en, el procesador o procesadores de hardware, o CPU. Este es el “lenguaje” mediante el cual el sistema y el software de aplicación se comunican con los procesadores de hardware. Cada instrucción nativa es un código discreto que se reconoce por la arquitectura de procesamiento y que puede especificar registros particulares para funciones aritméticas, de direccionamiento o de control; ubicaciones o compensaciones de memoria particulares; y modos de direccionamiento particulares utilizados para interpretar operandos. Las operaciones más complejas se construyen combinando estas instrucciones nativas sencillas, que se ejecutan secuencialmente, o según lo indiquen de otro modo las instrucciones de flujo de control.

La memoria 109 de sistema incluye múltiples memorias 111 intermedias en tesela, cada una configurada para almacenar una parte de los datos de imagen adquiridos de la celda de flujo durante un único ciclo de adquisición de imágenes. Las memorias intermedias en este ejemplo se denominan así porque están configuradas para contener los datos de imagen de una única tesela. Tal como se explica más detalladamente a continuación, una tesela es una región de una celda de flujo que puede capturarse en una única imagen tomada durante un único ciclo de adquisición de imágenes. Las memorias 111 intermedias están destinadas a almacenar datos de imagen durante un período de tiempo más largo que las memorias intermedias de procesador 107. En determinadas realizaciones, las memorias 111 intermedias en tesela almacenan datos de imagen durante al menos dos ciclos de adquisición de imágenes. Aunque esta solicitud describe memorias intermedias que almacenan datos de una tesela de una celda de flujo, las realizaciones descritas no se limitan a memorias intermedias que almacenan esta cantidad de datos. A menos que se indique lo contrario o esté claro a partir del contexto, se entiende que las referencias a “memorias intermedias en tesela” incluyen cualquier tipo de memoria intermedia que almacena datos de imagen de una parte de una celda de flujo, cuyos datos de imagen se procesan como una unidad tal como se describe en la presente memoria.

Para realizar llamadas de base, el uno o más procesadores 105 actúan sobre los datos proporcionados desde la memoria 109 de sistema y los datos almacenados en las memorias 107 intermedias de procesador. Normalmente, se realiza una única llamada de base para un único sitio durante un único ciclo de adquisición de imágenes.

5 Tal como se muestra, el uno o más procesadores 105 y la memoria 109 principal comparten datos bidireccionalmente. Adicionalmente, el uno o más procesadores 105 reciben datos de imagen del sistema 103 de adquisición de imágenes. En determinadas realizaciones, el sistema 103 de adquisición de imágenes obtiene datos de la celda 101 de flujo excitando los sitios de secuenciación en la celda 101 de flujo y recibiendo señales ópticas de esos sitios. En determinadas realizaciones, la señal recibida por el sistema 103 de adquisición de imágenes es una señal de fluorescencia creada cuando el sistema 103 ilumina la celda 101 de flujo con luz a longitudes de onda apropiadas. En tales realizaciones, la señal de fluorescencia se proporciona como valores de intensidad para una pluralidad de colores.

15 A lo largo de esta descripción se utiliza el concepto de un ciclo. Un único ciclo de secuenciación implica leer un único nucleótido de cada uno de uno o más sitios capturados en una imagen. La lectura se denomina hacer una llamada de base. En diversas realizaciones descritas en la presente memoria, un único ciclo computacional, desde la perspectiva del procesador o procesadores y la memoria, realiza tanto llamada de base como captura de imagen pero para diferentes nucleótidos, con la llamada de base rezagando la captura de imagen en la secuencia de nucleótidos que se lee o se llama. Por ejemplo, en un único ciclo computacional, el uno o más procesadores realizan llamada de base para un nucleótido en el ciclo de secuenciación n y al mismo tiempo realizan captura de imagen para el nucleótido en el ciclo de secuenciación $n + 1$. Por lo tanto, en un único ciclo computacional, el secuenciador (a) almacena y procesa datos de imagen no modificados para nucleótidos en el ciclo de secuenciación $n + 1$ y (b) hace una llamada de base para nucleótidos en el ciclo de secuenciación n . El uso de las memorias intermedias de procesador y las memorias intermedias de tesela en este procesamiento de ciclo por ciclo se describirá con más detalle a continuación.

25 Fase general

En un sitio particular en una celda de flujo u otro sustrato, se analizan juntas múltiples copias de una molécula de ácido nucleico, que tienen todas la misma secuencia (posiblemente con variaciones limitadas introducidas involuntariamente por el procesamiento de la muestra). Se usan copias suficientes para garantizar que se produzca una señal suficiente para permitir una llamada de base fiable. La colección de moléculas de ácido nucleico en un sitio se denomina agrupación. En algunos casos, una agrupación no secuenciada contiene solo moléculas de ácido nucleico monocatenario.

35 El ajuste de fase representa un artefacto no deseado que surge de la secuenciación de múltiples moléculas de ácido nucleico dentro de una agrupación. El ajuste de fase es la velocidad a la que las señales, tales como la fluorescencia de moléculas individuales dentro de una agrupación, pierden sincronización entre sí. A menudo, el término ajuste de fase se reserva para la señal de contaminación procedente de algunas moléculas que se encuentran detrás, y el término preajuste de fase se usa para la señal de contaminación procedente de otras moléculas que van por delante. El ajuste de fase y el preajuste de fase juntos describen el rendimiento del aparato de secuenciación y la química.

45 Los números bajos son mejores. Los valores de 0,10/0,10 significan que el 0,10 % de las moléculas en una agrupación se quedan atrás y el 0,10 % se queda por delante en cada ciclo de llamada de base. En otras palabras, el 0,20 % de la señal verdadera se pierde cada ciclo y, por lo tanto, contribuirá al ruido. Otro ejemplo, 0,20/0,20 significa que el 0,4 % de la señal verdadera se pierde por ciclo, en cuyo caso después de 250 ciclos (sin corrección) el ruido sería igual a la señal.

50 Un componente de análisis en tiempo real de un secuenciador puede determinar el ajuste de fase y el preajuste de fase para aplicar el nivel correcto de corrección de ajuste de fase a medida que avanza la secuenciación. Esto funciona empujando de manera artificial la señal hacia o fuera de cada canal de secuenciador basándose en las llamadas de base antes o después del ciclo actual.

55 Anteriormente, el ajuste de fase y el preajuste de fase se estimaba en un número definido de ciclos (por ejemplo, los primeros 12 ciclos de cada lectura) y luego se aplicaban a todos los ciclos posteriores. Algunos secuenciadores recientes emplean un algoritmo llamado corrección de ajuste de fase empírica para optimizar la corrección de ajuste de fase en cada ciclo probando un intervalo de correcciones y seleccionando el que da como resultado la mayor limpieza (pureza de señal). Aunque la corrección de ajuste de fase empírica proporciona un rendimiento mejorado, requiere mayores recursos computacionales.

60 En los secuenciadores convencionales, cada base tiene un color de colorante fluorescente único; por ejemplo, verde para timina, rojo para citosina, azul para guanina y amarillo para adenina. Para capturar información para la llamada de base, un secuenciador de cuatro canales toma cuatro imágenes de una tesela u otra parte de una celda de flujo. Algunos secuenciadores ahora tienen solo dos canales y, por lo tanto, solo toman dos imágenes de la misma parte de la celda de flujo. Un secuenciador de dos canales utiliza una mezcla de colorantes para cada base y utiliza filtros rojo y verde para las dos imágenes. En un ejemplo de un secuenciador de dos canales, las agrupaciones observadas en

imágenes rojas o verdes se interpretan como bases de C y T, respectivamente. Las agrupaciones observadas en imágenes tanto rojas como verdes se marcan como bases A, mientras que las agrupaciones no marcadas se identifican como bases G.

La figura 2 ilustra el ajuste de fase durante la secuenciación de una agrupación de nucleótidos que tiene la secuencia. . . ACGTAAG. . . Como se ilustra, durante el ciclo de llamada de base para la primera G, el 98,4 % de la señal de fluorescencia se origina a partir de secuencias que generan actualmente la señal para G, mientras que el 1,5 % de la señal de fluorescencia se origina a partir de secuencias que producen actualmente señal para la base C anterior, y el 1,1 % de la señal de fluorescencia se origina a partir de secuencias que producen actualmente señal para la siguiente base T. La contribución de la señal para la base C anterior proviene del ajuste de fase y la contribución de la señal de la siguiente base T proviene del preajuste de fase.

La corrección de ajuste de fase para esta llamada de base de G se refleja en el gráfico en el lado derecho de la figura 2. Tal como se muestra para un secuenciador de dos canales, la señal de fluorescencia puede representarse en un gráfico bidimensional, representando la señal de intensidad máxima en un “eje verde” T, representando la intensidad máxima en un “eje rojo” C, representando la intensidad máxima a medio camino entre los ejes A, y representando la intensidad mínima en ambos ejes G. Sin error de ajuste de fase, la señal para G debe tener intensidad cero en ambos ejes rojo y verde. En cambio, con el error de ajuste de fase comentado, la señal de fluorescencia tiene alguna contribución de intensidad en ambos ejes verde y rojo. En este ejemplo, la corrección de preajuste de fase reduce la intensidad de la señal a cero en el eje verde y la corrección de ajuste de fase reduce la intensidad de la señal a cero en el eje rojo. Pueden realizarse correcciones similares en las llamadas de base para las bases T, C y A.

Teselas y celdas de flujo

Tal como se ha explicado, una celda de flujo contiene múltiples sitios donde se recoge la información de secuenciación. En determinadas realizaciones, cada sitio de una celda de flujo contiene una agrupación de ácidos nucleicos monocatenarios que comparten la misma secuencia. Una única imagen usada en la secuenciación en tiempo real puede contener millones de tales agrupaciones. Una celda de flujo típica es tan grande que requiere cientos o incluso miles de imágenes independientes para cubrir su área completa. En determinadas realizaciones, el procesador y la memoria asociada empleados para el análisis en tiempo real procesan todas estas imágenes actualmente para realizar llamadas de base para un único ciclo. En algunas implementaciones, el procesador y la memoria procesan simultáneamente todas las imágenes adquiridas en dos o más celdas de flujo durante un único ciclo de llamada de base. La figura 3 representa esquemáticamente una arquitectura de celda de flujo utilizada en algunos secuenciadores de Illumina, Inc. En el ejemplo representado, el secuenciador realiza llamadas de base simultáneas en dos celdas de flujo, celda 1 de flujo y la celda 2 de flujo. En determinadas realizaciones, cada celda de flujo tiene sitios de secuenciación en cada una de dos superficies, una superficie superior y una superficie inferior. En tales casos, el secuenciador toma imágenes de las superficies tanto superior como inferior durante cada ciclo de llamada de base. Como se representa en la figura 3, cada superficie de celda de flujo incluye cuatro carriles, L1, L2, L3 y L4; por supuesto, son posibles otros números. Cada carril de cada superficie puede tener múltiples subdivisiones denominadas franjas. Cada franja se divide a su vez en múltiples teselas. Por ejemplo, puede haber aproximadamente 120 teselas por franja. Considerando dos celdas de flujo, cada una con dos superficies, teniendo cada superficie cuatro carriles, teniendo cada carril seis franjas y teniendo cada franja 120 teselas, deben analizarse varias miles de teselas de datos por ciclo. En diversas realizaciones, cada imagen de tesela (u otra imagen de una parte de una celda de flujo) se procesa por un único subproceso de procesador. En determinadas realizaciones, un secuenciador que emplea una celda de flujo que tiene la arquitectura representada en la figura 3 procesa 8000 o más teselas de datos en cada ciclo de llamada de base. En tales casos, la lógica de procesamiento en tiempo real emplearía 8000 o más subprocesos de procesador en cada ciclo de llamada de base.

Los datos de una única tesela capturados durante un único ciclo pueden almacenarse en la memoria como una matriz, representando cada entrada en la matriz un valor de color para cada canal de una única agrupación en la tesela. En la figura 4 se representa una matriz para una disposición de dos canales. Como ejemplo, un detector de intensidad de color puede generar recuentos de señal entre aproximadamente 400 y 1500 para cada canal. Una memoria intermedia en tesela en la memoria de sistema está configurada para almacenar toda la información en la matriz, en otras palabras, los valores de color de todas las agrupaciones en una tesela en un único ciclo de llamada de base. Una memoria intermedia de procesador puede configurarse de manera similar para almacenar toda la información en la matriz.

Procedimiento de ajuste de fase

Una carga de memoria significativa del análisis en tiempo real de los datos de secuencia proviene del requisito en la corrección de ajuste de fase de que se deben guardar dos o tres ciclos de intensidades de agrupación para cada tesela para toda la longitud de la ejecución. En un dispositivo Illumina HiSeqX con una celda de flujo de 700 nm, esto ocupa 73 Gigabytes de memoria. Esta carga es lo suficientemente grande como para que la mayoría de los datos (en esta plataforma) se guarden en caché en un disco duro de estado sólido.

Como se explicó, la corrección de ajuste de fase ajusta los valores de intensidad de una imagen para abordar la secuenciación en fase de algunos soportes de ácido nucleico en una agrupación. La corrección de ajuste de fase logra esto comenzando con los valores medidos de intensidad del color de agrupación (u otras señales medidas con el método de secuenciación) para un ciclo de llamada de base actual y añadiendo o restando un valor de corrección usando los valores de intensidad medidos del ciclo de llamada de base anterior y/o usando valores de intensidad medidos del ciclo de llamada de base posterior. En diversas implementaciones, un valor de intensidad con corrección de ajuste de fase para realizar una llamada de base aplica una expresión como se muestra en la parte inferior de la figura 5. Tal como se muestra en ella, los valores de intensidad con corrección de ajuste de fase para un ciclo de llamada de base actual en una imagen equivalen a los valores de intensidad medidos para el ciclo de llamada de base actual menos el producto de un primer coeficiente y los valores de intensidad medidos en el ciclo de llamada de base inmediatamente anterior y menos el producto de un segundo coeficiente y valores de intensidad medidos en el ciclo de llamada de base inmediatamente sucesivo:

$$\text{Intensidad corregida} = -a \cdot I_{n-1} + I_n - b \cdot I_{n+1}$$

donde I_{n-1} , I_n , e I_{n+1} son los valores de intensidad de las agrupaciones en una tesela en el ciclo de llamada de base inmediatamente anterior, en el ciclo de llamada de base actual, y en el ciclo de llamada de base inmediatamente posterior, respectivamente. Los coeficientes a y b son los coeficientes de ajuste de fase y preajuste de fase (a veces denominados ponderaciones), respectivamente. Estos pueden calcularse de nuevo para cada ciclo de llamada de base de una tesela.

Volviendo a la figura 2, el valor de intensidad medido para la tercera base en la secuencia representada (para una única agrupación en una imagen) se muestra como un punto en el gráfico en el lado derecho de la figura 2. La corrección de preajuste de fase de este valor de intensidad medido se refleja por la flecha vertical desde el valor de intensidad medido hasta el eje horizontal. En la expresión para los valores de intensidad con corrección de ajuste de fase, esta corrección de preajuste de fase se representa por el producto del coeficiente b y el valor de intensidad medido para el siguiente ciclo de llamada de base posterior. Además, el valor de intensidad medido se corrige mediante una corrección de ajuste de fase representada por la flecha horizontal en el gráfico. Esta corrección de ajuste de fase se implementa restando del valor de intensidad medido, el producto de un coeficiente a y el valor de intensidad medido para el ciclo de llamada de base inmediatamente anterior. Los coeficientes a y b pueden determinarse mediante numerosos métodos, pero en muchas implementaciones, se calculan de nuevo para cada ciclo de llamada de base. Una descripción de métodos para determinar los coeficientes que se utilizarán en la corrección de ajuste de fase se describe en la solicitud de patente internacional con número de publicación WO 2015/084985 de Belitz y col. y publicada el 11 de junio de 2015.

En determinadas realizaciones, el algoritmo de ajuste de fase determina los coeficientes de ajuste de fase empíricamente maximizando la limpieza acumulativa (o métrica similar) de los datos de intensidad de agrupación durante un ciclo de llamada de base. Una implementación del algoritmo itera sobre todos o muchos coeficientes de ajuste de fase y determina cuáles dan los mejores resultados. Por ejemplo, el algoritmo de ajuste de fase puede optimizar a y b en cada ciclo utilizando una búsqueda de patrón que emplea una función de coste que cuenta el número de agrupaciones que fallan en un filtro de limpieza. Por lo tanto, a y b se seleccionan para maximizar la calidad de los datos.

En algunas realizaciones, los coeficientes de ajuste de fase se determinan como un análisis continuo a lo largo de una ejecución de secuenciación (por ejemplo, durante la generación de una lectura). Como resultado de este enfoque, una estimación de ajuste de fase inexacta realizada durante ciclos tempranos no afectará negativamente a ciclos posteriores.

Algunos métodos determinan la limpieza de un valor de intensidad de agrupación en función de las distancias relativas a los centroides gaussianos para los otros valores de intensidad de agrupación determinados para el mismo ciclo de llamada de base. Los centroides se alinean idealmente con las ubicaciones esperadas de las intensidades de A, T, C y G para dos canales (véase la figura 2), suponiendo que se use un sistema de dos canales. En determinadas realizaciones, la limpieza puede calcularse utilizando la expresión:

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

donde $D1$ es la distancia al centroide gaussiano más cercano, y $D2$ es la distancia al siguiente centroide más cercano. Utilizando este enfoque, cuando se maximiza la limpieza media (calidad) de los valores de intensidad, se eligen los valores correctos de a y b . Una vez que se identifican estos valores, entonces puede aplicarse una corrección a todos los valores de agrupación y puede producirse directamente la llamada de base. Los métodos de ajuste de distribuciones gaussianas a un conjunto de datos de dos canales se describen en la solicitud de patente internacional que tiene el número de publicación WO 2015/084985.

En algunas realizaciones, se calcula una corrección de ajuste de fase en casi todos los ciclos durante una ejecución de secuenciación. En algunas realizaciones, se calcula una corrección de ajuste de fase en cada ciclo durante una ejecución de secuenciación. En algunas realizaciones, se calcula una corrección de ajuste de fase independiente para diferentes ubicaciones de una superficie obtenida como imagen en el mismo ciclo. Por ejemplo, en algunas realizaciones, se calcula una corrección de ajuste de fase independiente para cada carril individual de una superficie obtenida como imagen, tal como un carril de celda de flujo individual. En algunas

realizaciones, se calcula una corrección de ajuste de fase independiente para cada subconjunto de un carril, tal como una franja de obtención de imágenes dentro de un carril de celda de flujo. En algunas realizaciones, se calcula una corrección de ajuste de fase independiente para cada imagen individual, tal como, por ejemplo, cada tesela. En determinadas realizaciones, se calcula una corrección de ajuste de fase independiente para cada tesela en cada ciclo.

A medida que las lecturas son más largas, los términos de orden superior pueden volverse más importantes en la 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 tal como se define por lo siguiente:

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

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

La figura 5 representa esquemáticamente una configuración de procesamiento y metodología para realizar corrección de ajuste de fase en tiempo real. En la realización representada, un procesador 502 crea un nuevo subproceso 503 de procesamiento cuando se solicita al procesador que realice llamadas desde agrupaciones en una imagen, por ejemplo, una imagen de una tesela. Puede generarse un nuevo subproceso para cada ciclo de llamada de base para cada tesela. En la realización representada, el procesador 502 pone a disposición una única memoria 505 intermedia de procesador para cada ciclo de llamada de base de una tesela (y el subproceso de procesamiento designado). La memoria intermedia de procesador almacena temporalmente valores de intensidad que el procesador manipula computacionalmente para llevar a cabo la corrección de ajuste de fase para un ciclo de llamada de base actual n . En la realización representada, el procesador interactúa con una memoria 507 de sistema que contiene tres memorias intermedias, cada una para almacenar datos de imagen capturados para un ciclo de llamada de base particular. En el caso de la arquitectura de celda de flujo representada en la figura 3, cada memoria intermedia almacena datos de imagen para las agrupaciones de una única tesela; por lo tanto, las memorias intermedias se denominan memorias intermedias en tesela. Por supuesto, para otras arquitecturas de celdas de flujo y/o sistemas de adquisición de imágenes, las memorias intermedias pueden almacenar más o menos datos de agrupación. Por conveniencia, la memoria descriptiva se referirá a memorias intermedias en tesela. Cada memoria intermedia en tesela almacena datos para una única tesela (u otra parte de una celda de flujo) capturada durante un único ciclo de llamada de base. Los datos de imagen pueden proporcionarse como una matriz de datos tal como se muestra en la figura 4.

Como se representa, la memoria 507 de sistema incluye una memoria 509 intermedia en tesela que almacena temporalmente valores de intensidad para el ciclo de llamada de base inmediatamente anterior (en comparación con el ciclo de llamada de base actual manejado por el procesador), una memoria 511 intermedia en tesela que almacena valores de intensidad medidos para el ciclo de llamada de base actual, y una memoria 513 intermedia en tesela que almacena valores de intensidad para el ciclo de llamada de base inmediatamente posterior. De nuevo, cada una de las memorias 509, 511 y 513 intermedias en tesela contiene datos medidos de una única tesela para un único ciclo de llamada de base n .

Tal como se muestra, el subproceso 503 hace uso de los valores de intensidad en cada una de las memorias 509, 511 y 513 intermedias en tesela durante un único ciclo de llamada de base. Los valores de intensidad se cargan sucesivamente en la memoria 505 intermedia de procesador y se manipulan para implementar la expresión de corrección de ajuste de fase presentada en la parte inferior de la figura 5. Una vez que se completa el procedimiento de llamada de base como se representa en la configuración de procesador y memoria de la figura 5, la memoria intermedia de procesador contiene valores de intensidad ajustados utilizados para hacer una llamada de base con corrección de ajuste de fase.

La figura 6 presenta un diagrama de flujo de un procedimiento de llamada de base que puede emplear la configuración de procesador y memoria representada en la figura 5. Tal como se muestra en la figura 6, un procedimiento 601 inicia un nuevo ciclo de llamada de base creando un subproceso de procesador y asignando una memoria intermedia de procesador a ese subproceso. Véase el bloque 603 de procedimiento. Posteriormente, el procesador extrae datos de intensidad de una imagen de una tesela de celda de flujo (u otra parte apropiada de la celda de flujo) tomada simultáneamente con el ciclo de procesamiento actual. En la implementación representada, la imagen capturada y los valores de intensidad asociados son los valores de intensidad primarios para el siguiente ciclo de llamada de base posterior, no el ciclo de llamada de base actual (la iteración de procesamiento actual). En otras palabras, el ciclo de procesamiento actual realiza una llamada de base para los datos de imagen recogidos en un ciclo de procesamiento inmediatamente anterior. Por lo tanto, como se representa en un bloque 605 de procedimiento del procedimiento 601, los valores de intensidad extraídos reciben la referencia I_{n+1} , w , donde n representa el ciclo de llamada de base actual. Dicho de otra manera, un ciclo de procesamiento tanto (i) llama bases para el ciclo de llamada de base n , como (ii) captura datos de imagen para el ciclo de llamada de base $n+1$.

Los datos de intensidad recién extraídos, que pueden proporcionarse en forma de una matriz como se representa en la figura 4, se almacenan en una memoria intermedia en tesela disponible en la memoria de sistema (por ejemplo, la memoria 513 intermedia en tesela). En determinadas realizaciones, esta memoria intermedia en tesela es aquella que almacenó datos de intensidad que se usaron previamente pero ya no son necesarios para la llamada de base.

En el ciclo de procesamiento actual, el procedimiento 601 también recupera datos de intensidad almacenados durante un ciclo computacional previo al actual ciclo computacional. Véase el bloque 607 de procedimiento. Los datos de intensidad recuperados son para el ciclo de llamada de base actual y se les da la referencia I_n . Los datos de intensidad recuperados se obtienen a partir de una memoria intermedia en tesela apropiada, tal como la memoria 511 intermedia en tesela de la memoria de sistema, tal como se muestra en la figura 5.

Además, el procedimiento 601 recupera datos de intensidad que se almacenaron dos ciclos antes del ciclo de llamada de base actual. Véase el bloque 609 de procedimiento. Como ejemplo, con referencia a la figura 5, tales datos de intensidad pueden obtenerse de una memoria 509 intermedia en tesela de la memoria de sistema. El conjunto de valores de intensidad recuperados en la operación 609 se identifica I_{n-1} .

Si bien se muestra que las operaciones 605, 607 y 609 se producen secuencialmente, este orden de operaciones es flexible y el procedimiento puede implementarse de manera que cualquier orden sea aceptable, siempre que esté de acuerdo con la llamada de base que incorpora la corrección de ajuste de fase.

Al recuperar los valores de intensidad para el ciclo de llamada de base actual (bloque 607 de procedimiento) y los valores de intensidad para el ciclo de llamada de base inmediatamente anterior (bloque 609 de procesamiento), el procesador tiene disponibles todos los valores de intensidad que necesita para realizar una corrección de ajuste de fase. Esto lo realiza determinando primero la ponderación de corrección de preajuste de fase b y la ponderación de corrección de ajuste de fase a para el ciclo de llamada de base actual. Véase el bloque 611 de procedimiento, que ilustra que esto puede lograrse mediante el uso de los valores de intensidad extraídos para el siguiente ciclo de llamada de base posterior junto con los valores de intensidad para los ciclos de llamada de base actual e inmediatamente anterior. A continuación, usando las ponderaciones de corrección de ajuste de fase y preajuste de fase, el procesador calcula valores de intensidad con corrección de ajuste de fase para el ciclo de llamada de base actual como se representa en el bloque 613 de procedimiento. Los valores corregidos son para las agrupaciones en la tesela en consideración. El cálculo puede emplear la expresión representada en el bloque 613. Usando los valores de intensidad con corrección de ajuste de fase, el procesador realiza llamadas para el ciclo de llamada de base actual como se representa en el bloque 615 de procedimiento.

En este punto, se completa el procesamiento para el ciclo de llamada de base actual y puede ejecutarse la siguiente iteración de llamada de base. La decisión de si realizar otro ciclo de llamada de base se representa en un bloque 617 que determina si hay ningún nucleótido adicional para secuenciar en las agrupaciones de la tesela en consideración. Si no hay ninguno, el procedimiento se completa como se representa en el bloque 619. Si los hay, el control de procedimiento pasa a un bloque 621 de procedimiento donde el procesador incrementa un recuento de ciclos. Esto indexa de manera eficaz los valores de intensidad para el ciclo de llamada de base actual I_n a los valores de la intensidad para el ciclo de llamada de base inmediatamente anterior I_{n-1} . Al mismo tiempo, los valores de intensidad para el ciclo de llamada de base inmediatamente siguiente (I_{n+1}) se convierten en los valores de intensidad para el nuevo ciclo de llamada de base actual (I_n). Estos incrementos se realizan con respecto a los índices aplicados a los datos de intensidad almacenados en las memorias intermedias en tesela.

Procedimiento de ajuste de fase (memoria principal reducida)

El enfoque de las figuras 5 y 6 puede funcionar bien siempre que el secuenciador y su sistema de análisis en tiempo real asociado no estén limitados en memoria. Sin embargo, dada la cantidad de datos que deben procesarse en determinados secuenciadores modernos, tales como los empleados para realizar la secuenciación del genoma completo, es posible que no se disponga de memoria insuficiente, en particular a un coste comercialmente viable. Por lo tanto, almacenar tres veces la cantidad de datos requeridos para obtener una imagen completa de la celda de flujo (o celdas de flujo) durante un ciclo de llamada de base puede presentar un grave cuello de botella.

Un algoritmo de ajuste de fase tal como el representado en las figuras 5 y 6 es una contribución importante al análisis en tiempo real, ya que mejora significativamente los resultados de secuenciación, particularmente en muestras no convencionales, por ejemplo, muestras de baja diversidad. Sin embargo, la carga de memoria impuesta se vuelve mayor a medida que crece el rendimiento de los sistemas de secuenciación de próxima generación. Las siguientes realizaciones reducen la carga de memoria usando ponderaciones de ajuste de fase aprendidas a partir de datos con corrección de ajuste de fase parcial previa. Las ponderaciones de ajuste de fase y preajuste de fase pueden aprenderse de forma independiente y aún proporcionar resultados de secuenciación de alta calidad. En algunos ejemplos, el requisito de memoria principal es menor que dos veces la capacidad requerida para almacenar los datos contenidos en el número total de teselas en dos celdas de flujo.

En determinadas realizaciones, se ajusta la configuración de procesador y memoria para la llamada de base con corrección de ajuste de fase para reducir los requisitos de memoria de sistema. En la figura 7 se muestra un ejemplo de cómo funciona esto. Los valores de intensidad se corrigen como se describió anteriormente, por ejemplo, se calculan las ponderaciones de ajuste de fase y preajuste de fase y se aplican a los ciclos inmediatamente anterior e inmediatamente posterior. Sin embargo, en el ejemplo de la figura 7, la memoria 707 de sistema solo emplea dos memorias intermedias en tesela para la corrección de ajuste de fase: la memoria 709 intermedia en tesela y la memoria 711 intermedia en tesela. En este ejemplo, un procesador 702 emplea un subproceso 703 de procesamiento que, al contrario del ejemplo de la figura 5, tiene dos memorias intermedias de procesador asociadas: una memoria 705 intermedia de procesador para almacenar y operar sobre los valores de intensidad recuperados de la memoria 707 y una memoria 706 intermedia de procesador para almacenar y usar los valores de intensidad de imagen recién capturados I_{n+1} . En el ejemplo representado, las memorias intermedias de procesador se asignan en memoria principal, pero esto no siempre es necesario. En algunas realizaciones, las memorias intermedias de procesador se asignan en una memoria física diferente o incluso en el chip de procesador.

El reemplazo de las memorias intermedias en tesela por las memorias intermedias de procesador reduce eficazmente los requisitos de memoria total. Usando múltiples procesadores y/o procesamiento de subprocesos múltiples, unos pocos procesadores manejan muchas teselas. Como ejemplo, el número de teselas en un sistema puede ser del orden de 1000-2000, mientras que el número de procesadores que manejan todas estas teselas es de aproximadamente veinte. En teoría, dicho sistema puede realizar una reducción de memoria del orden de 50x. En algunas implementaciones, la reducción es del orden de 20x.

En esta implementación, los valores de intensidad capturados de las imágenes de tesela en el ciclo de procesamiento actual (I_{n+1}) se almacenan localmente en el procesador y se usan para calcular las ponderaciones de ajuste de fase y preajuste de fase y, posteriormente, realizan una llamada de base. En algunas implementaciones, solo después de que se completa este procedimiento, los valores de intensidad capturados más recientemente (I_{n+1}) se almacenan en una memoria intermedia en tesela en la memoria 707 de sistema.

En algunas realizaciones, un procesador y una memoria de sistema están configurados como se representa en la figura 8. Al igual que con la configuración de procesador/memoria en la figura 7, un procesador 802 emplea subprocesos 808 de procesamiento, cada uno asociado con dos memorias intermedias de procesador: una memoria 805 intermedia de procesador para almacenar temporalmente valores de intensidad de una memoria 807 de sistema (memoria 811 intermedia en tesela), y una memoria 806 intermedia de procesador para almacenar temporalmente valores de intensidad capturados durante el ciclo de procesamiento actual (I_{n+1}). Con el fin de permitir que esta configuración funcione de manera eficiente y eficaz, los valores de intensidad almacenados en la memoria 811 intermedia en tesela deben tener corrección de ajuste de fase parcial. A continuación se describen ejemplos de mecanismos para lograr esto. La memoria 705 intermedia de procesador en la figura 7 y la memoria 805 intermedia de procesador en la figura 8 cargan intensidades desde la memoria principal y luego manipulan esas intensidades para generar las intensidades corregidas que se emplean para la llamada de base. En el ejemplo representado, las memorias intermedias de procesador se asignan en memoria principal, pero esto no siempre es necesario. En algunas realizaciones, las memorias intermedias de procesador se asignan en una memoria física diferente o incluso en el chip de procesador.

La figura 9 presenta una vista de alto nivel de un procedimiento 901 que puede emplearse con la configuración de procesador y memoria de la figura 8 y, en algunas implementaciones, la figura 7. Como se ilustra en la figura 9, los ciclos de procesamiento primero y segundo emplean información insuficiente para realizar una corrección de ajuste de fase completa en las agrupaciones tomadas como imagen en una tesela. Sin embargo, la fase no es un problema significativo en los primeros ciclos.

Para realizar la corrección de ajuste de fase completa, el secuenciador requiere tres ciclos consecutivos de datos de imagen. En el primer ciclo de procesamiento, el secuenciador no realiza una llamada de base; simplemente almacena los datos de intensidad para el siguiente procesamiento, es decir, el ciclo en el que se realiza la primera llamada de base.

Como se representa, el procedimiento 901 comienza en un bloque 903 de procedimiento donde se crea un subproceso para el primer ciclo de procesamiento. Las instrucciones de este subproceso dirigen la extracción de datos de intensidad de una imagen de las agrupaciones durante el primer ciclo de secuenciación (I_1), es decir, el ciclo durante el cual se leen los primeros nucleótidos de las agrupaciones. Véase el bloque 905 de procedimiento. Los datos de imagen se almacenan en una memoria intermedia en tesela en la memoria de sistema. En este punto, el primer ciclo de procesamiento se completa eficazmente.

El procedimiento continúa en un bloque 907 de procedimiento donde se crea un nuevo subproceso en preparación para el segundo ciclo de procesamiento. En este procedimiento, se asignan las memorias intermedias del primer y segundo procesador para el segundo ciclo de procesamiento. Véase el bloque 907. En conjunto, los bloques 907, 909, 911, 913, 915, 917, 919, 921 y 923 de procedimiento se realizan durante el segundo ciclo de procesamiento, que se ejecuta usando las memorias intermedias de subproceso y procesador generadas en el bloque 907 de procedimiento.

Como se representa, el procesador extrae los datos de intensidad de la imagen para el siguiente ciclo de llamada de base (I_2) y almacena estos datos en una primera memoria intermedia de procesador. Véase el bloque 909 de procedimiento. A continuación, durante el segundo ciclo de procesamiento, el procesador recupera los datos de intensidad almacenados en la memoria intermedia en tesela durante el primer ciclo de procesamiento, los datos de intensidad que son para el ciclo de llamada de base actual (I_1). Véase el bloque 911. Mediante el uso de los datos de intensidad recogidos durante los ciclos de procesamiento primero y segundo, el procesador puede calcular una ponderación de preajuste de fase b para el ciclo de llamada de base actual (es decir, las primeras llamadas de base en las lecturas). Véase el bloque 913 de procedimiento. Con los valores de intensidad para los primeros dos ciclos y la ponderación de preajuste de fase, el procesador calcula valores de datos de intensidad corregidos para el segundo ciclo de llamada de base (I_2). Los valores de datos de intensidad corregidos pueden almacenarse en la segunda memoria intermedia de procesador. Véase el bloque 915 de procedimiento. A continuación, el procesador realiza las llamadas de base para el segundo ciclo de llamada de base usando los valores de datos de intensidad corregidos obtenidos en el bloque 915. Véase el bloque 917 de procedimiento.

En este punto, el procedimiento de secuenciación está listo para comenzar a prepararse para el siguiente ciclo de llamada de base. Se inicia en un bloque 919 de procedimiento determinando una ponderación de corrección de ajuste de fase a usando los siguientes (o segundos) datos de intensidad de ciclo de llamada de base (I_2) y los datos de ciclo de llamada de base actual (I_1), que se almacenaron en la memoria intermedia en tesela. Usando la ponderación de corrección de ajuste de fase a , el procesador calcula a continuación los valores de datos de intensidad con corrección de ajuste de fase (pero sin corrección de preajuste de fase) a partir de los datos de intensidad no corregidos actualmente (I_2) extraídos durante este segundo ciclo de procesamiento y los valores de datos de intensidad para el primer ciclo de procesamiento (I_1) según la expresión presentada en el bloque 921 de procedimiento. Esto da como resultado una matriz de valores de intensidad parcialmente corregidos ($I_{2(\text{parcialmente corregida})}$) para el segundo ciclo de llamada de base. El secuenciador tendrá que esperar el siguiente ciclo de procesamiento antes de realizar la corrección de preajuste de fase. Sin embargo, en este punto se completa gran parte del cálculo y los datos de matriz para una única imagen pueden almacenarse en una memoria intermedia en tesela para su uso en el siguiente ciclo de llamada de base. Para este fin, el procesador almacena los datos de intensidad con corrección de ajuste de fase (pero sin corrección de preajuste de fase) en la memoria intermedia en tesela (de tal manera que $I_{2(\text{parcialmente corregida})}$ reemplaza a I_1 en la memoria intermedia en tesela). Véase el bloque 923 de procedimiento.

En este punto, se completan los ciclos de procesamiento primero y segundo y se realizan llamadas de base para el primer ciclo de llamada de base, que es el segundo ciclo de procesamiento. Pueden realizarse ciclos de llamada de base posteriores con corrección de ajuste de fase completa tal como se describe en la figura 10. Véase el bloque 925 de procedimiento.

La figura 10 representa una secuencia de operaciones que puede realizarse durante un ciclo de procesamiento que lleva a cabo la llamada de base con corrección de ajuste de fase completa. Un ciclo de este tipo puede realizarse en los ciclos de procesamiento tercero y posteriores cuando se secuencian agrupaciones de una tesela. En determinadas realizaciones, la secuencia de operaciones representada en la figura 10 corresponde al bloque 925 de procedimiento de la figura 9.

Como se representa, el procedimiento comienza mediante la asignación de un subproceso y las memorias intermedias de procesador primera y segunda. Véase el bloque 1003 de procedimiento. A continuación, el procesador extrae valores de datos de intensidad de una imagen para el siguiente ciclo de llamada de base (I_{n+1}) y almacena esos valores en una primera memoria intermedia de procesador. Véase el bloque 1005 de procedimiento. Al mismo tiempo, el procesador recupera los valores de datos de intensidad parcialmente corregidos que se almacenaron durante el ciclo de llamada de base anterior (como un ejemplo no limitativo, $I_{2(\text{parcialmente corregida})}$ en la realización de la figura 9, o $I_n - a(I_{n-1})$). Estos valores representan ahora los valores de intensidad para el ciclo de llamada de base actual (I_n). Anteriormente se almacenaron en la memoria intermedia en tesela de la memoria de sistema y ahora se recuperan de la misma. Véase el bloque 1007 de procedimiento. Con los valores de datos de intensidad parcialmente corregidos para el ciclo de llamada de base actual, que se corrigió en ajuste de fase, el procesador solo necesita realizar la corrección de preajuste de fase para completar la corrección de los datos de intensidad y realizar las llamadas de base necesarias para el ciclo de llamada de base actual. Para este fin, el procesador determina la ponderación de corrección de preajuste de fase b para el ciclo de llamada de base actual. Esto lo realiza utilizando datos de intensidad extraídos que acaba de recuperar de los datos de imagen, para el siguiente ciclo (I_{n+1}), junto con los datos de intensidad parcialmente corregidos con anterioridad para el ciclo de llamada de base actual. Debe recordarse que esto corrigió parcialmente los datos que acababan de recuperarse de la memoria intermedia en tesela. Los datos de intensidad parcialmente corregidos pueden estar representados por la expresión $I_n - a(I_{n-1})$. Véase el bloque 1009 de procedimiento.

Con la ponderación de corrección de preajuste de fase b calculada para el ciclo de llamada de base actual, el procesador tiene todo lo que necesita para calcular una matriz de datos de intensidad con corrección de ajuste de fase completa para el ciclo de llamada de base actual (I_n). El cálculo se realiza como se representa en el bloque 1009 de procedimiento. Los valores de datos de intensidad corregidos completamente resultantes se almacenan en la segunda memoria intermedia de procesador. Véase el bloque 1011 de procedimiento. A continuación, el procesador realiza las llamadas de base para el ciclo de llamada de base actual usando los valores de datos de

intensidad corregidos almacenados en la segunda memoria intermedia de procesador. Véase el bloque 1013 de procedimiento.

El ciclo de procesamiento actual puede comenzar a prepararse para el siguiente ciclo de llamada de base que se ejecutará durante el siguiente ciclo de procesamiento. En la realización representada, el procesador determina la ponderación de corrección de ajuste de fase *a* para el siguiente ciclo de llamada de base utilizando datos de intensidad disponibles para el ciclo de llamada de base actual. Véase el bloque 1015 de procedimiento. Debe recordarse que los datos de intensidad del siguiente ciclo de llamada de base se extrajeron y se almacenaron en la primera memoria intermedia de procesador en la operación 1005 de procedimiento. Los valores de intensidad parcialmente corregidos para el ciclo de llamada de base actual se recuperaron de la memoria intermedia en tesela para hacer las llamadas de base actuales. Los mismos valores de intensidad parcialmente corregidos se utilizan ahora para calcular la ponderación de corrección de ajuste de fase *a* para el siguiente ciclo de llamada de base. Con la ponderación de corrección de ajuste de fase para el siguiente ciclo de llamada de base ahora calculado, el procesador calcula valores de datos de intensidad corregidos (pero sin corrección de preajuste de fase) como se representa en el bloque 1017 de procedimiento. El procesador almacena entonces estos valores de datos de intensidad con corrección de ajuste de fase para el siguiente ciclo de llamada de base en la memoria intermedia en tesela. Véase el bloque 1019 de procedimiento.

Antes de esta invención, se suponía que la precisión de la llamada de base se vería afectada al aprender las ponderaciones de preajuste a partir de las intensidades con corrección de fase. Sin embargo, los resultados en la presente memoria muestran que se producen pocos o ningún resultado con imprecisión. En algunas implementaciones, los datos de imagen se comprimen (por ejemplo, compresión con pérdidas) e incluso se comprimen los datos con corrección de fase parcial. En ambos casos, se ha demostrado que la compresión podría realizarse sin pérdida de precisión. Como ejemplo, sin compresión, una implementación utiliza dos memorias intermedias flotantes para cada tesela (una memoria intermedia flotante tiene un tamaño de 4 bytes). Con compresión, una implementación utiliza una memoria intermedia de un único byte, consiguiendo de este modo 4 veces menos memoria.

En este punto, el ciclo de procesamiento actual se completa eficazmente, por lo que el procesador determina si hay más ciclos que deben realizarse para secuenciar las agrupaciones de la tesela actual. Véase el bloque 1021 de decisión. Si no es necesario leer más bases de las agrupaciones, el procedimiento se completa y no se realizan ciclos de procesamiento adicionales. Sin embargo, si se requieren uno o más ciclos de secuenciación adicionales, el control del procedimiento se dirige a un bloque 1023 de procedimiento donde el procesador incrementa el ciclo actual en cuyo punto los valores de datos de intensidad parcialmente corregidos almacenados en la memoria intermedia en tesela se convierten en actuales; es decir, se convierten en los valores para el nuevo ciclo de llamada de base. El control de procedimiento vuelve entonces al bloque 1003 de procedimiento donde comienza el siguiente ciclo de procesamiento.

Ejemplos

Como se ha explicado, determinadas realizaciones reducen la carga de memoria usando ponderaciones de ajuste de fase aprendidas a partir de los datos con corrección de ajuste de fase parcial previa. Sin embargo, no estaba claro que las ponderaciones de ajuste de fase y preajuste de fase puedan aprenderse de forma independiente y aún proporcionar resultados de secuenciación de alta calidad. El ejemplo presentado en la figura 11 establece que sí pueden.

Tal como se muestra, se realizaron dos comparaciones, cada una usando un procedimiento de referencia (por ejemplo, un procedimiento de las figuras 5 y 6) y un nuevo procedimiento que se optimizó para reducir los requisitos de memoria principal (por ejemplo, un procedimiento de las figuras 8 y 10). En cada comparación, se empleó el mismo secuenciador y la misma muestra. Específicamente, se convirtió un instrumento Illumina HiSeqX para usar 2 productos químicos colorantes. Las imágenes de salida del secuenciador se guardaron y los dos algoritmos de fase se probaron en las mismas imágenes de secuenciación, proporcionando una prueba completamente controlada. Las “agrupaciones PF” indican el rendimiento proporcionado por el secuenciador; el % alineado indica el número de grupos que se alinearon satisfactoriamente con el genoma de referencia, y el “% de tasa de error” indica la tasa de error media de las secuencias llamadas por el software en comparación con el genoma de referencia.

Los resultados de secuenciación demuestran que el algoritmo de ajuste de fase eficiente en memoria es comparable al algoritmo de referencia. En este ejemplo, el procedimiento eficiente en memoria produjo un aumento de aproximadamente el 3 % en la tasa de error, que se compensa por una reducción en la memoria principal (estimada de desde 420 Gigabytes hasta 340 Gigabytes en algunas implementaciones).

Métodos de secuenciación

Como se indicó anteriormente, la descripción se refiere a la secuenciación de muestras de ácido nucleico. Puede usarse cualquiera de una serie de tecnologías de secuenciación que usan uno o más canales de información para la llamada de base, particularmente canales ópticos. Las técnicas particularmente aplicables son aquellas en las que los ácidos nucleicos se unen en ubicaciones fijas en una matriz (por ejemplo, como agrupamiento) y donde se obtienen imágenes de la matriz de manera repetida. Son particularmente aplicables las realizaciones en las

que se obtienen imágenes en diferentes canales de color, por ejemplo, coincidiendo con diferentes marcadores utilizados para distinguir un tipo de base de nucleótido de otro. En algunas realizaciones, el proceso para determinar la secuencia de nucleótidos de un ácido nucleico diana puede ser un proceso automatizado. Determinadas realizaciones incluyen técnicas de secuenciación por síntesis ("SBS"). Aunque en este caso se hace hincapié en las técnicas de secuenciación por síntesis, pueden emplearse otras tecnologías de secuenciación.

En muchas implementaciones, las técnicas de SBS implican la extensión enzimática de una cadena de ácido nucleico en formación mediante la adición iterativa de nucleótidos contra una cadena molde. En los métodos tradicionales de SBS, se puede proporcionar un monómero de un solo nucleótido a un nucleótido diana en presencia de una polimerasa en cada administración. Sin embargo, en los métodos descritos en la presente descripción, se puede proporcionar más de un tipo de monómero nucleotídico a un ácido nucleico diana en presencia de una polimerasa en una administración.

La SBS puede utilizar monómeros de nucleótidos que tienen un resto terminador o aquellos que carecen de restos terminadores. Los métodos que utilizan monómeros de nucleótidos que carecen de terminadores incluyen, por ejemplo, pirosecuenciación y secuenciación utilizando nucleótidos marcados con γ -fosfato. En los métodos que utilizan monómeros de nucleótidos que carecen de terminadores, el número de nucleótidos añadidos en cada ciclo es generalmente variable y depende de la secuencia molde y del modo de administración de nucleótidos. Para las técnicas de SBS que utilizan monómeros de nucleótidos que tienen un resto terminador, el terminador puede ser efectivamente irreversible en las condiciones de secuenciación utilizadas como es el caso de la secuenciación tradicional de Sanger que utiliza didesoxinucleótidos, o el terminador puede ser reversible como es el caso de los métodos de secuenciación desarrollados por Solexa (ahora Illumina, Inc.).

Las técnicas de SBS pueden utilizar monómeros de nucleótidos que tienen un resto marcador o aquellos que carecen de un resto marcador. Por consiguiente, los acontecimientos de incorporación pueden detectarse basándose en una característica del marcador, tal como fluorescencia del marcador; una característica del monómero de nucleótido tal como el peso molecular o la carga; un subproducto de incorporación del nucleótido, tal como la liberación de pirofosfato; o similares. En las realizaciones, donde están presentes dos o más nucleótidos diferentes en un reactivo de secuenciación, los diferentes nucleótidos pueden distinguirse entre sí o, como alternativa, los dos o más marcadores diferentes pueden ser indistinguibles según las técnicas de detección que se utilizan. Por ejemplo, los diferentes nucleótidos presentes en un reactivo de secuenciación, pueden tener diferentes marcadores y se pueden distinguir usando ópticas apropiadas como se ilustra mediante los métodos de secuenciación desarrollados por Solexa (ahora Illumina, Inc.).

Algunas realizaciones incluyen técnicas de pirosecuenciación. La pirosecuenciación detecta la liberación de pirofosfato inorgánico (PPI) a medida que se incorporan nucleótidos particulares en la cadena en formación (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 estadounidense n.º 6.210.891; patente estadounidense US-6.258.568 y Pat. US-6.274.320). En la pirosecuenciación, el PPI liberado puede detectarse convirtiéndolo inmediatamente en adenosín trifosfato (ATP) por la ATP sulfurilasa, y el nivel de ATP generado se detecta mediante fotones producidos por luciferasa. Los ácidos nucleicos que se van a secuenciar se pueden unir a características en una matriz y se pueden obtener imágenes de la matriz para capturar las señales quimioluminiscentes que se producen debido a la incorporación de nucleótidos en las características de la matriz. Puede obtenerse una imagen después de que la matriz se trate 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 las características de la matriz que se detecten. Estas diferencias en la imagen reflejan el contenido de secuencia diferente 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 utilizando los métodos expuestos en la presente descripción. Por ejemplo, las imágenes obtenidas después del tratamiento de la matriz con cada tipo de nucleótido diferente pueden manejarse de la misma manera que se ilustra en la presente memoria para imágenes obtenidas de diferentes canales de detección para métodos de secuenciación basados en terminadores reversibles.]

En otro tipo ilustrativo de SBS, la secuenciación del ciclo se logra mediante la adición escalonada de nucleótidos terminadores reversibles que contienen, por ejemplo, un marcador de colorante escindible o fotoblanqueable como se describe, por ejemplo, en el documento WO 04/018497 y en la patente estadounidense n.º 7.057.026. Este enfoque está siendo comercializado por Solexa (ahora Illumina Inc.), y también se describe en los documentos WO 91/06678 y WO 07/123.744.. La disponibilidad de terminadores marcados con fluorescencia, en los que se puede invertir tanto la terminación como escindir el marcador fluorescente, facilita la secuenciación eficaz de la cyclic reversible termination (terminación cíclica reversible - CRT). Las polimerasas también pueden diseñarse conjuntamente para incorporar y extenderse eficazmente a partir de estos nucleótidos modificados.

En realizaciones de secuenciación basadas en terminadores reversibles, los marcadores pueden no inhibir sustancialmente la extensión en las condiciones de reacción de SBS. Sin embargo, los marcadores de detección

pueden eliminarse, por ejemplo, mediante escisión o degradación. Las imágenes pueden capturarse tras la incorporación de marcadores en características de ácidos nucleicos en matriz. En realizaciones particulares, cada ciclo implica la administración simultánea de cuatro tipos de nucleótidos diferentes a la matriz y cada tipo de nucleótido tiene un marcador espectralmente distinto. A continuación, se pueden obtener cuatro imágenes, cada una de las cuales utiliza un canal de detección que es selectivo para uno de los cuatro marcadores diferentes. Como alternativa, se pueden añadir secuencialmente diferentes tipos de nucleótidos y se puede obtener 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 dichos métodos de SBS con terminador reversible pueden almacenarse, procesarse y analizarse como se expone en la presente descripción. Después de la etapa de captura de imágenes, los marcadores se pueden eliminar y los restos terminadores reversibles se pueden eliminar para los ciclos posteriores de adición y detección de nucleótidos. La eliminación de los marcadores después de que se hayan detectado en un ciclo particular y antes de un ciclo posterior, puede proporcionar la ventaja de reducir la señal de fondo y la diafonía entre ciclos.

En realizaciones particulares, algunos o todos los monómeros de nucleótidos pueden incluir terminadores reversibles. En tales realizaciones, los terminadores reversibles/compuestos de flúor escindibles pueden incluir flúor unido al resto de ribosa mediante un enlace éster 3' (Metzker, *Genome Res.* 15:1767-1776 [2005]). Otros enfoques han separado la química del terminador de la escisión del marcador de fluorescencia (Ruparel y col., *Proc Natl Acad Sci EE. UU.* 102: 5932-7 [2005]). Ruparel y col. describieron el desarrollo de terminadores reversibles que usaban un pequeño grupo alilo 3' para bloquear la extensión, pero podían desbloquearse fácilmente mediante un tratamiento corto con un catalizador de paladio. El fluoróforo se unió a la base mediante un ligador fotoescindible que podía escindirse fácilmente mediante una exposición de 30 segundos a luz UV de longitud de onda larga. Por lo tanto, puede usarse o bien una reducción de disulfuro o bien fotoescisión como un ligador escindible. Otro enfoque para la terminación reversible es el uso de la terminación natural que se produce después de la colocación de un tinte voluminoso en un dNTP. La presencia de un tinte voluminoso cargado en el dNTP puede actuar como un terminador eficaz a través de un impedimento estérico y/o electrostático. La presencia de un evento de incorporación evita otras incorporaciones a menos que se elimine el tinte. La división del tinte elimina el flúor e invierte eficazmente la terminación. También se describen ejemplos de nucleótidos modificados en la patente estadounidense n.º 7.427.673 y en la patente estadounidense n.º 7.057.026.

Los sistemas y métodos de SBS ilustrativos adicionales que pueden utilizarse con los métodos y sistemas descritos en la presente memoria se describen en la publicación de solicitud de patente estadounidense n.º 2007/0166705, la publicación de solicitud de patente estadounidense n.º 2006/0188901, la patente estadounidense US-7.057.026, la publicación de solicitud de patente estadounidense n.º 2006/0240439, la publicación de solicitud de patente estadounidense n.º 2006/0281109, la publicación PCT n.º WO 05/065814, la publicación de solicitud de patente estadounidense n.º 2005/0100900, la publicación PCT n.º WO 06/064199, la publicación PCT n.º WO 07/010.251, la publicación de solicitud de patente estadounidense n.º 2012/0270305 y la publicación de solicitud de patente estadounidense n.º 2013/0260372.

Algunas realizaciones pueden utilizar la detección de cuatro nucleótidos diferentes usando menos de cuatro marcadores diferentes. Por ejemplo, la SBS puede realizarse utilizando métodos y sistemas descritos en los materiales incorporados de la publicación de solicitud de patente estadounidense n.º 2013/0079232. Como primer ejemplo, puede detectarse un par de tipos de nucleótidos en la misma longitud de onda, pero distinguirse en función de la diferencia de intensidad de un miembro del par en comparación con el otro, o en función de un cambio en un miembro del par (por ejemplo, mediante modificación química, modificación fotoquímica o modificación física) que hace que aparente aparezca o desaparezca una señal aparente en comparación con la señal detectada para el otro miembro del par. Como segundo ejemplo, pueden detectarse tres de los cuatro tipos de nucleótidos diferentes en condiciones particulares, mientras que un cuarto tipo de nucleótido carece de un marcador que sea detectable en esas condiciones, o se detecta mínimamente en esas condiciones (por ejemplo, detección mínima debido a la fluorescencia de fondo, etc.). La incorporación de los primeros tres tipos de nucleótidos en un ácido nucleico se puede determinar basándose en la presencia de sus respectivas señales y la incorporación del cuarto tipo de nucleótidos en el ácido nucleico se puede determinar basándose en la ausencia o detección mínima de cualquier señal. Como un tercer ejemplo, un tipo de nucleótidos puede incluir un(os) marcador(es) que se detecta(n) en dos canales diferentes, mientras que otros tipos de nucleótidos se detectan en no más de uno de los canales. Las tres configuraciones ilustrativas mencionadas anteriormente no se consideran mutuamente excluyentes y pueden usarse en varias 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 un marcador que se detecta en el primer canal cuando se excita con una primera longitud de onda de excitación), un segundo tipo de nucleótido que se detecta en un segundo canal (por ejemplo, dCTP que tiene un marcador que se detecta en el segundo canal cuando se excita con una segunda longitud de onda de excitación), un tercer tipo de nucleótido que se detecta tanto en el primer como en el segundo canal (por ejemplo, dTTP que tiene al menos un marcador que se detecta en ambos canales cuando se excita con la primera y/o segunda longitud de onda de excitación) y un cuarto tipo de nucleótido que carece de un marcador que no se detecta, o se detecta mínimamente, en ninguno de los canales (por ejemplo, dGTP sin marcador).

Además, como se describe en los materiales incorporados de la publicación de patente estadounidense n.º 2013/0079232, los datos de secuenciación pueden obtenerse utilizando un único canal. En los llamados enfoques

de secuenciación de un tinte, el primer tipo de nucleótido se marca pero el marcador se elimina después de que se genere la primera imagen, y el segundo tipo de nucleótido se marca solo después de que se genere una primera imagen. El tercer tipo de nucleótido conserva su marcador tanto en la primera como en la segunda imagen, y el cuarto tipo de nucleótido permanece sin marcar en ambas imágenes.

Algunas realizaciones pueden utilizar secuenciación mediante técnicas de ligamiento. Tales técnicas utilizan ADN ligasa para incorporar oligonucleótidos e identificar la incorporación de tales oligonucleótidos. Los oligonucleótidos tienen normalmente diferentes marcadores que se correlacionan con la identidad de un nucleótido particular en una secuencia con la que hibridan los oligonucleótidos. Al igual que con otros métodos de SBS, se pueden obtener imágenes tras el tratamiento de una matriz de características de ácidos nucleicos con los reactivos de secuenciación marcados. Cada imagen mostrará características de ácido nucleico que han incorporado marcadores de un tipo particular. Diferentes características estarán presentes o ausentes en las diferentes imágenes debido al contenido de secuencia diferente 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 ligamiento se pueden almacenar, procesar y analizar como se expone en la presente descripción. Sistemas y métodos de SBS ilustrativos que pueden utilizarse con los métodos y sistemas descritos en la presente memoria se describen en la patente estadounidense n.º 6.969.488, la patente estadounidense n.º 6.172.218, y la patente estadounidense n.º 6.306.597.

Algunas realizaciones pueden utilizar la secuenciación de nanoporos (Deamer, D. W. y 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 una proteína de membrana biológica, tal como la α -hemolisina. Cuando el ácido nucleico diana pasa a través del nanoporo, cada par de bases puede identificarse midiendo las fluctuaciones en la conductancia eléctrica del poro. (Patente estadounidense n.º 7.001.792; Soni, G. V. y 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]).

Los datos obtenidos de la secuenciación de nanoporos se pueden almacenar, procesar y analizar como se expone en la presente descripción. En particular, los datos se pueden tratar como una imagen según el tratamiento ilustrativo de imágenes ópticas y otras imágenes que se expone en la presente descripción.

Algunas realizaciones pueden utilizar métodos que suponen la monitorización en tiempo real de la actividad de la ADN polimerasa. Las incorporaciones de nucleótidos pueden detectarse mediante interacciones de transferencia de energía por resonancia de fluorescencia (FRET) entre una polimerasa portadora de fluoróforo y nucleótidos marcados con γ -fosfato como se describe, por ejemplo, en la patente estadounidense n.º 7.329.492 y la patente estadounidense n.º 7.211.414 o las incorporaciones de nucleótidos pueden detectarse con guías de onda en modo cero como se describe, por ejemplo, en la patente estadounidense n.º 7.315.019 y usando análogos de nucleótidos fluorescentes y polimerasas diseñadas mediante ingeniería como se describe, por ejemplo, en la patente estadounidense n.º 7.405.281 y la publicación de solicitud de patente estadounidense n.º 2008/0108082. La iluminación se puede restringir a un volumen a escala de zeptolitros alrededor de una polimerasa anclada a la superficie, de modo que se pueda observar la incorporación de nucleótidos marcados con fluorescencia con un fondo bajo (Levene, M. J. y col. "Zero-mode waveguides for single-molecule analysis at high concentrations." Science 299, 682-686 (2003); Lundquist, P. M. y col. "Parallel confocal detection of single molecules in real time." Opt. Lett. 33, 1026-1028 (2008); Korlach, J. y col. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures." Proc. Natl. Acad. Sci. EE. UU. 105, 1176-1181 [2008]). Las imágenes obtenidas de dichos métodos se pueden almacenar, procesar y analizar tal como se expone en la presente descripción.

Algunas realizaciones de SBS incluyen la detección de un protón liberado tras la incorporación de un nucleótido en un producto de extensión. Por ejemplo, la secuenciación basada en la detección de protones liberados puede usar un detector eléctrico y técnicas asociadas que están disponibles comercialmente en Ion Torrent (Guilford, CT, una subsidiaria de Life Technologies) o métodos y sistemas de secuenciación descritos en los documentos US 2009/0026082 A1; US 2009/0127589 A1; US 2010/0137143 A1; o US 2010/0282617 A1. Los métodos expuestos en la presente descripción para amplificar ácidos nucleicos diana usando exclusión cinética pueden aplicarse fácilmente a sustratos usados para detectar protones. Más específicamente, los métodos expuestos en la presente descripción pueden usarse para producir poblaciones clonales de amplicones que se usan para detectar protones.

Los métodos de SBS anteriores se pueden llevar a cabo ventajosamente en formatos multiplexados de modo que se manipulen simultáneamente múltiples ácidos nucleicos diana diferentes. En realizaciones particulares, se pueden tratar diferentes ácidos nucleicos diana en un recipiente de reacción común o en una superficie de un sustrato particular. Esto permite la administración conveniente de reactivos de secuenciación, la eliminación de reactivos que

no han reaccionado y la detección de eventos de incorporación de una manera multiplexada. En realizaciones que usan ácidos nucleicos diana unidos a la superficie, los ácidos nucleicos diana pueden estar en un formato de matriz. En un formato de matriz, los ácidos nucleicos diana pueden unirse normalmente a una superficie de una manera espacialmente distinguible. Los ácidos nucleicos diana pueden unirse mediante unión covalente directa, unión a una perla u otra partícula o unión a una polimerasa u otra molécula que esté unida a la superficie. La matriz puede incluir una única copia de un ácido nucleico diana en cada sitio (también denominado como una característica) o pueden estar presentes múltiples copias que tienen la misma secuencia en cada sitio o característica. Pueden producirse múltiples copias mediante métodos de amplificación tales como la amplificación en puente o la PCR en emulsión.

Los métodos expuestos en la presente descripción pueden usar matrices que tienen características en cualquiera de una variedad de densidades que incluyen, por ejemplo, al menos aproximadamente 10 características/cm², 100 características/cm², 500 características/cm², 1000 características/cm², 5000 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 superior.

Una ventaja de los métodos expuestos en la presente memoria es que proporcionan una detección rápida y eficaz de una pluralidad de ácidos nucleicos diana en paralelo. Por consiguiente, la presente descripción proporciona sistemas integrados capaces de preparar y detectar ácidos nucleicos usando técnicas conocidas en la técnica tales como las ilustradas anteriormente. Por tanto, un sistema integrado de la presente descripción puede incluir componentes fluidicos capaces de administrar reactivos de amplificación y/o reactivos de secuenciación a uno o más fragmentos de ADN inmovilizados, comprendiendo el sistema componentes tales como bombas, válvulas, depósitos, líneas fluidicas y similares. Una celda de flujo puede configurarse y/o usarse en un sistema integrado para la detección de ácidos nucleicos diana. Por ejemplo, se describen celdas de flujo ilustrativas en los documentos US 2010/0111768 A1 y US con n.º de serie 13/273.666. Como se ilustra en las celdas de flujo, uno o más de los componentes fluidicos de un sistema integrado se pueden usar para un método de amplificación y para un método de detección. Tomando una realización de secuenciación de ácidos nucleicos como ejemplo, se pueden usar uno o más de los componentes fluidicos de un sistema integrado para un método de amplificación expuesto en la presente descripción y para la administración de reactivos de secuenciación en un método de secuenciación como los ilustrados anteriormente. Como alternativa, un sistema integrado puede incluir sistemas fluidicos separados para llevar a cabo métodos de amplificación y para llevar a cabo métodos de detección. Los ejemplos de sistemas de secuenciación integrados que son capaces de crear ácidos nucleicos amplificados y también de determinar la secuencia de los ácidos nucleicos incluyen, sin limitación, la plataforma MiSeq™ (Illumina, Inc., San Diego, CA) y los dispositivos descritos en el documento estadounidense con número de serie 13/273.666.

En algunas realizaciones de los métodos descritos en la presente memoria, las etiquetas de secuencia mapeadas comprenden lecturas de secuencia de aproximadamente 20 pb, aproximadamente 25 pb, aproximadamente 30 pb, aproximadamente 35 pb, aproximadamente 40 pb, aproximadamente 45 pb, aproximadamente 50 pb, aproximadamente 55 pb, aproximadamente 60 pb, aproximadamente 65 pb, aproximadamente 70 pb, aproximadamente 75 pb, aproximadamente 80 pb, aproximadamente 85 pb, aproximadamente 90 pb, aproximadamente 95 pb, aproximadamente 100 pb, aproximadamente 110 pb, aproximadamente 120 pb, aproximadamente 130 pb, aproximadamente 140 pb, aproximadamente 150 pb, aproximadamente 200 pb, aproximadamente 250 pb, aproximadamente 300 pb, aproximadamente 350 pb, aproximadamente 400 pb, aproximadamente 450 pb o aproximadamente 500 pb. En algunos casos, se emplean lecturas de un solo extremo de más de 500 pb para lecturas de más de aproximadamente 1000 pb cuando se generan lecturas de extremo emparejadas. El mapeo de las etiquetas de secuencia se logra comparando la secuencia de la etiqueta con la secuencia de la referencia para determinar el origen cromosómico de la molécula de ácido nucleico secuenciada, y no se necesita información de secuencia genética específica. Puede permitirse un pequeño grado de emparejamientos erróneos (0-2 emparejamientos erróneos por lectura) para explicar polimorfismos minoritarios que pueden existir entre el genoma de referencia y los genomas en la muestra mixta.

Sistemas y aparato para el análisis en tiempo real de datos de secuenciación

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

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

En diversas realizaciones, los datos o información empleada en los métodos y aparatos descritos se proporciona en un formato electrónico. Tales datos o información pueden incluir lecturas derivadas de una muestra de ácido nucleico, recuentos o densidades de tales etiquetas que se alinean con regiones particulares de una secuencia de referencia (por ejemplo, que se alinean con un cromosoma o segmento cromosómico), distancias de separación entre lecturas o fragmentos adyacentes, distribuciones de tales distancias de separación, diagnósticos y similares. Como se utiliza en la presente memoria, los datos u otra información proporcionada en formato electrónico está disponible para su almacenamiento en una máquina y transmisión entre máquinas. Convencionalmente, los datos en formato electrónico se proporcionan digitalmente y pueden almacenarse como bits y/o bytes en diversas estructuras de datos, listas, bases de datos, etc. Los datos pueden incorporarse electrónicamente, ópticamente, etc.

Una realización proporciona un producto de programa informático para determinar los coeficientes de ajuste de fase y preajuste de fase, así como valores de magnitud con corrección de ajuste de fase y llamadas de base asociadas. El producto informático puede contener instrucciones para realizar uno cualquiera o más de los métodos descritos anteriormente para el ajuste de fase y la llamadas de base. Como se ha explicado, el producto informático puede incluir un medio legible por ordenador no transitorio y/o tangible que tiene una lógica compilable o ejecutable por ordenador (por ejemplo, instrucciones) grabada en el mismo para permitir que un procesador alinee lecturas, identifique fragmentos y/o islas a partir de lecturas alineadas, identifique alelos, incluyendo alelos de indel, de polimorfismos heterocigotos, partes de fase de cromosomas y cromosomas y genomas de haplotipos. En un ejemplo, el producto informático incluye (1) un medio legible por ordenador que tiene una lógica compilable o ejecutable por ordenador (por ejemplo, instrucciones) almacenada en el mismo para permitir que un procesador realice la corrección de ajuste de fase en datos de magnitud (por ejemplo, datos de intensidad de color de dos o más canales) en muestras de ácido nucleico; (2) lógica asistida por ordenador para realizar llamadas de base de las muestras de ácido nucleico; y (3) un procedimiento de salida para generar una salida que caracteriza las muestras de ácido nucleico.

Debe entenderse que no es práctico, o incluso posible en la mayoría de los casos, que un ser humano sin ayuda realice las operaciones computacionales de los métodos descritos en la presente memoria. Por ejemplo, generar coeficientes de ajuste de fase incluso para una única tesela durante un único ciclo de llamada de base podría requerir años de esfuerzo sin la ayuda de un aparato informático. Por supuesto, el problema se agrava porque la secuenciación NGS fiable generalmente requiere la corrección de ajuste de fase y la llamada de base para al menos miles o incluso millones de lecturas.

Los métodos descritos en la presente memoria pueden realizarse usando un sistema para secuenciar muestras de ácido nucleico. El sistema puede incluir: (a) un secuenciador para recibir ácidos nucleicos de la muestra de prueba que proporciona información de secuencia de ácidos nucleicos de la muestra; (b) un procesador; y (c) uno o más medios de almacenamiento legibles por ordenador que tienen almacenadas en los mismos instrucciones para su ejecución en el procesador para evaluar los datos del secuenciador. Los medios de almacenamiento legibles por ordenador también pueden almacenar datos de magnitud con corrección de ajuste de fase parcial de las agrupaciones en una celda de flujo.

En algunas realizaciones, los métodos se instruyen por un medio legible por ordenador que tiene almacenadas en el mismo instrucciones legibles por ordenador para llevar a cabo un método para determinar la fase de una secuencia. Por lo tanto, una realización proporciona un producto de programa informático que incluye uno o más medios de almacenamiento no transitorios legibles por ordenador que tienen almacenadas en el mismo instrucciones ejecutables por ordenador que, cuando se ejecutan por uno o más procesadores de un sistema informático, hacen que el sistema informático implemente un método para secuenciar una muestra de ADN. El método incluye: (a) obtener datos que representan una imagen (por ejemplo, la imagen en sí) de un sustrato que comprende una pluralidad de sitios donde se leen las bases de ácidos nucleicos; (b) obtener valores de color (u otros valores que representan bases/nucleótidos individuales) de la pluralidad de sitios a partir de la imagen del sustrato; (c) almacenar los valores de color en una memoria intermedia de procesador; (d) recuperar valores de color con corrección de fase parcial de la pluralidad de sitios para un ciclo de llamada de base, donde los valores de color con corrección de fase parcial se almacenaron en la memoria del secuenciador durante un ciclo de llamada de base inmediatamente anterior; (e) determinar una corrección de preajuste de fase a partir de (i) los valores de

color con corrección de ajuste de fase parcial almacenados durante el ciclo de llamada de base inmediatamente anterior, y (ii) los valores de color almacenados en la memoria intermedia de procesador; y (f) determinar valores de color corregidos a partir de (i) los valores de color en la memoria intermedia de procesador, (ii) los valores con corrección de fase parcial almacenados durante el ciclo inmediatamente anterior, y (iii) la corrección de preajuste de fase.

La secuencia u otros datos pueden introducirse en un ordenador o almacenarse en un medio legible por ordenador, ya sea directa o indirectamente. En diversas realizaciones, un sistema informático está integrado o acoplado directamente a un dispositivo de secuenciación que lee y/o analiza secuencias de ácidos nucleicos de muestras. Las secuencias u otra información de tales herramientas se proporcionan al sistema informático (o simplemente en hardware de procesamiento integrado) a través de una interfaz de transmisión de datos. Además, el dispositivo de memoria puede almacenar lecturas, información de calidad de llamada de base, información de coeficientes de ajuste de fase, etc. La memoria también puede almacenar diversas rutinas y/o programas para analizar y presentar los datos de secuencia. Tales programas/rutinas pueden incluir programas para realizar análisis estadísticos, etc.

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

En algunas realizaciones, los métodos también incluyen recoger datos con respecto a una pluralidad de secuencias de polinucleótido (por ejemplo, lecturas) y enviar los datos a un ordenador u otro sistema informático. Por ejemplo, el ordenador puede estar conectado a equipos de laboratorio, por ejemplo, un aparato de recogida de muestras, un aparato de amplificación de polinucleótidos, un aparato de secuenciación de nucleótidos. Los datos recogidos o almacenados pueden transmitirse desde el ordenador hasta una ubicación remota, por ejemplo, a través de una red local o una red de área amplia tal como internet. En la ubicación remota, pueden realizarse diversas operaciones en los datos transmitidos.

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

Conclusión

Las realizaciones descritas deben considerarse en todos los aspectos solo como ilustrativas y no restrictivas. Por lo tanto, el alcance de la descripción queda indicado por las reivindicaciones adjuntas más que por la descripción anterior.

REIVINDICACIONES

1. Un método para determinar valores de color corregidos a partir de datos de imagen adquiridos, durante un ciclo de llamada de base, por un secuenciador de ácidos nucleicos que comprende un sistema de adquisición de imágenes, uno o más procesadores, y una memoria que tiene preferiblemente una capacidad de almacenamiento de aproximadamente 512 Gigabytes o menos, realizándose el método en el uno o más procesadores y comprendiendo:
 - (a) obtener una imagen de un sustrato que comprende una pluralidad de sitios donde se leen las bases de ácidos nucleicos, en donde los sitios muestran colores que representan tipos de bases de ácidos nucleicos;
 - (b) medir valores de color de la pluralidad de sitios a partir de la imagen del sustrato;
 - (c) almacenar los valores de color en una memoria intermedia de procesador del uno o más procesadores del secuenciador;
 - (d) recuperar valores de color con corrección de fase de la pluralidad de sitios, en donde los valores de color con corrección de fase se almacenaron en la memoria del secuenciador durante un ciclo de llamada de base inmediatamente anterior
 - (e) determinar una corrección de preajuste de fase a partir de
 - los valores de color con corrección de fase almacenados durante el ciclo de llamada de base inmediatamente anterior, y
 - los valores de color almacenados en la memoria intermedia de procesador; y
 - (f) determinar los valores de color corregidos a partir de
 - los valores de color en la memoria intermedia de procesador,
 - los valores con corrección de fase almacenados durante el ciclo inmediatamente anterior, y
 - la corrección de preajuste de fase.
2. El método según la reivindicación 1, que comprende además usar los valores de color corregidos para realizar llamadas de base para la pluralidad de sitios; y/o
 - en donde la corrección de preajuste de fase comprende una ponderación y en donde determinar los valores de color corregidos comprende multiplicar la ponderación por los valores de color de la pluralidad de sitios medidos a partir de la imagen del sustrato.
3. El método según cualquiera de las reivindicaciones anteriores, que comprende además determinar una corrección de ajuste de fase para el ciclo de llamada de base inmediatamente posterior.
4. El método según la reivindicación 3, en donde determinar la corrección de ajuste de fase para el ciclo de llamada de base inmediatamente posterior comprende analizar
 - los valores de color con corrección de fase almacenados en la memoria del secuenciador, y
 - los valores de color almacenados en la memoria intermedia de procesador.
5. El método según la reivindicación 3, que comprende además:
 - producir valores de color con corrección de fase para el ciclo de llamada de base inmediatamente posterior aplicando la corrección de ajuste de fase a valores de color de la pluralidad de sitios almacenados en la memoria del secuenciador y
 - almacenar los valores de color con corrección de fase para el ciclo de llamada de base inmediatamente posterior en la memoria del secuenciador opcionalmente en donde:
 - producir los valores de color con corrección de fase para el ciclo de llamada de base inmediatamente posterior comprende además sumar
 - los valores de color con corrección de ajuste de fase de la pluralidad de sitios, y
 - los valores de color de la pluralidad de sitios de la imagen del sustrato medido en (b).
6. El método según cualquiera de las reivindicaciones anteriores, en donde el método se realiza en tiempo real durante la adquisición de lecturas de secuencia por el secuenciador de ácidos nucleicos.
7. El método según cualquiera de las reivindicaciones anteriores, en donde:
 - (i) los valores de color se determinan a partir de solo dos canales del secuenciador; o
 - (ii) los valores de color se obtienen a partir de cuatro canales del secuenciador.

- 5 8. El método según cualquiera de las reivindicaciones anteriores, que comprende además asignar la memoria intermedia de procesador y una segunda memoria intermedia de procesador, en donde la segunda memoria intermedia de procesador se usa para determinar los valores de color corregidos en (f).
9. Un secuenciador de ácidos nucleicos que comprende:
 - 10 un sistema de adquisición de imágenes;
una memoria que tiene preferiblemente una capacidad de almacenamiento de aproximadamente 512 Gigabytes o menos; y
uno o más procesadores diseñados o configurados para realizar el método según una cualquiera de las reivindicaciones 1 a 8.
- 15 10. El secuenciador de ácidos nucleicos según la reivindicación 9, en donde la memoria se divide en una pluralidad de memorias intermedias en tesela, cada una designada para almacenar datos que representan una única imagen de una tesela en el sustrato.
- 20 11. El secuenciador de ácidos nucleicos según la reivindicación 9 o 10, que comprende además un sistema para sintetizar ácidos nucleicos en la pluralidad de sitios.
- 25 12. El secuenciador de ácidos nucleicos según cualquiera de las reivindicaciones 9 a 11, en donde el sustrato comprende una celda de flujo, en donde la celda de flujo se divide lógicamente en teselas, y en donde cada tesela representa una región de la celda de flujo que comprende un subconjunto de sitios, subconjunto que se captura en una única imagen del sistema de adquisición de imágenes.
- 30 13. El secuenciador de ácidos nucleicos según la reivindicación 12, en donde, en la operación (d), los valores de color con corrección de fase se almacenaron en memorias intermedias en tesela de la memoria del secuenciador, y en donde las memorias intermedias en tesela se designan para almacenar datos que representan imágenes de teselas individuales sobre el sustrato.
- 35 14. El secuenciador de ácidos nucleicos según la reivindicación 12, en donde el uno o más procesadores están diseñados o configurados además para realizar las siguientes etapas:
 - antes de la operación (a), proporcionar reactivos a la celda de flujo y permitir que los reactivos interaccionen con sitios para mostrar los colores que representan tipos de bases de ácidos nucleicos durante el ciclo de llamada de base.
- 40 15. El secuenciador de ácidos nucleicos según la reivindicación 14, en donde el uno o más procesadores están diseñados o configurados además para realizar las siguientes etapas:
 - después de la operación (f):
 - 45 proporcionar reactivos nuevos a la celda de flujo y permitir que los reactivos nuevos interaccionen con los sitios para mostrar colores que representan tipos de bases de ácidos nucleicos para un siguiente ciclo de llamada de base; y
repetir las operaciones (a)-(e) para el siguiente ciclo de llamada de base; comprendiendo además opcionalmente crear un primer subproceso de procesador para realizar las operaciones (a)-(f) para el ciclo de llamada de base, y crear un segundo subproceso de procesador para realizar las operaciones (a)-(f) para el siguiente ciclo de llamada de base.

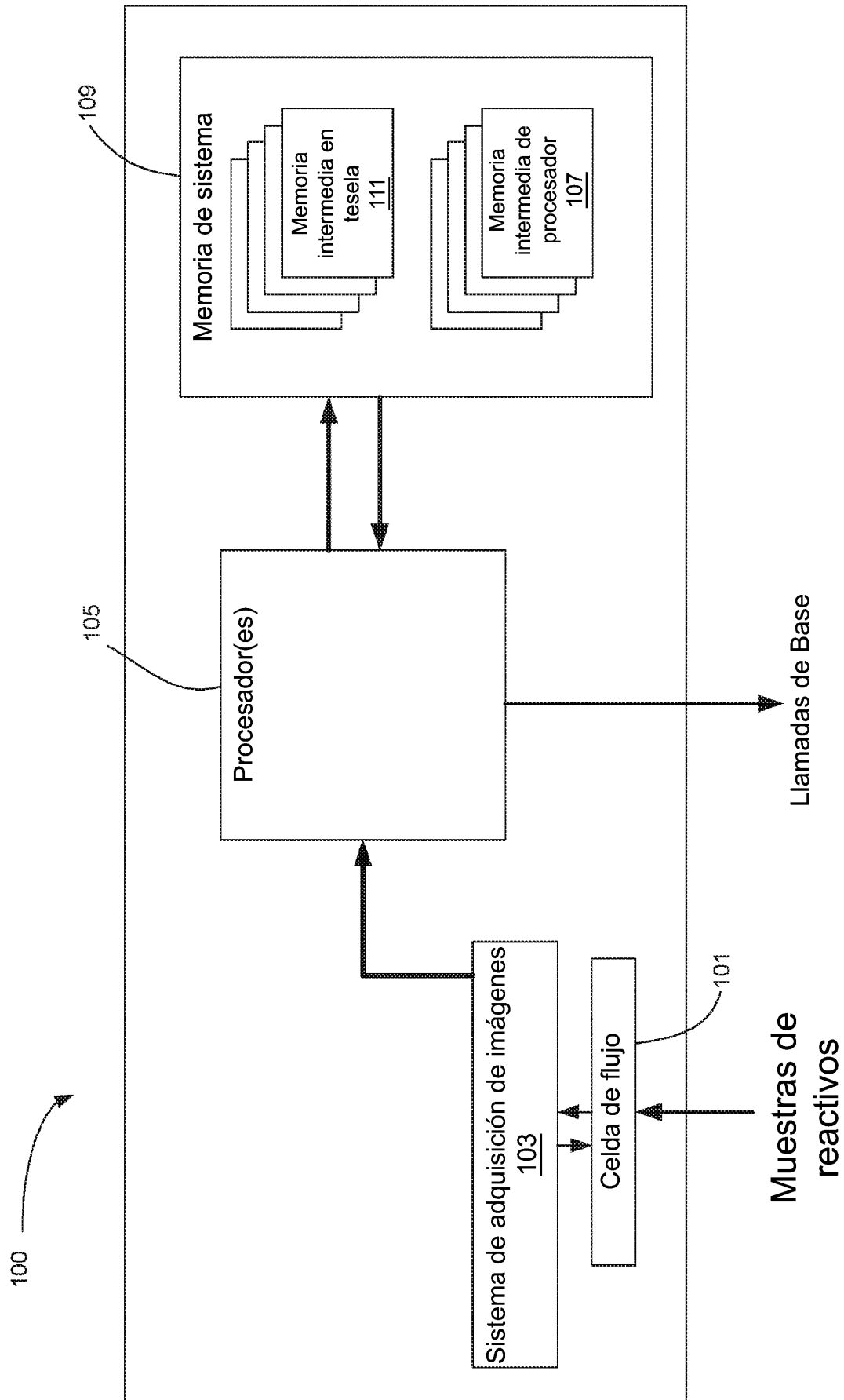


Figura 1

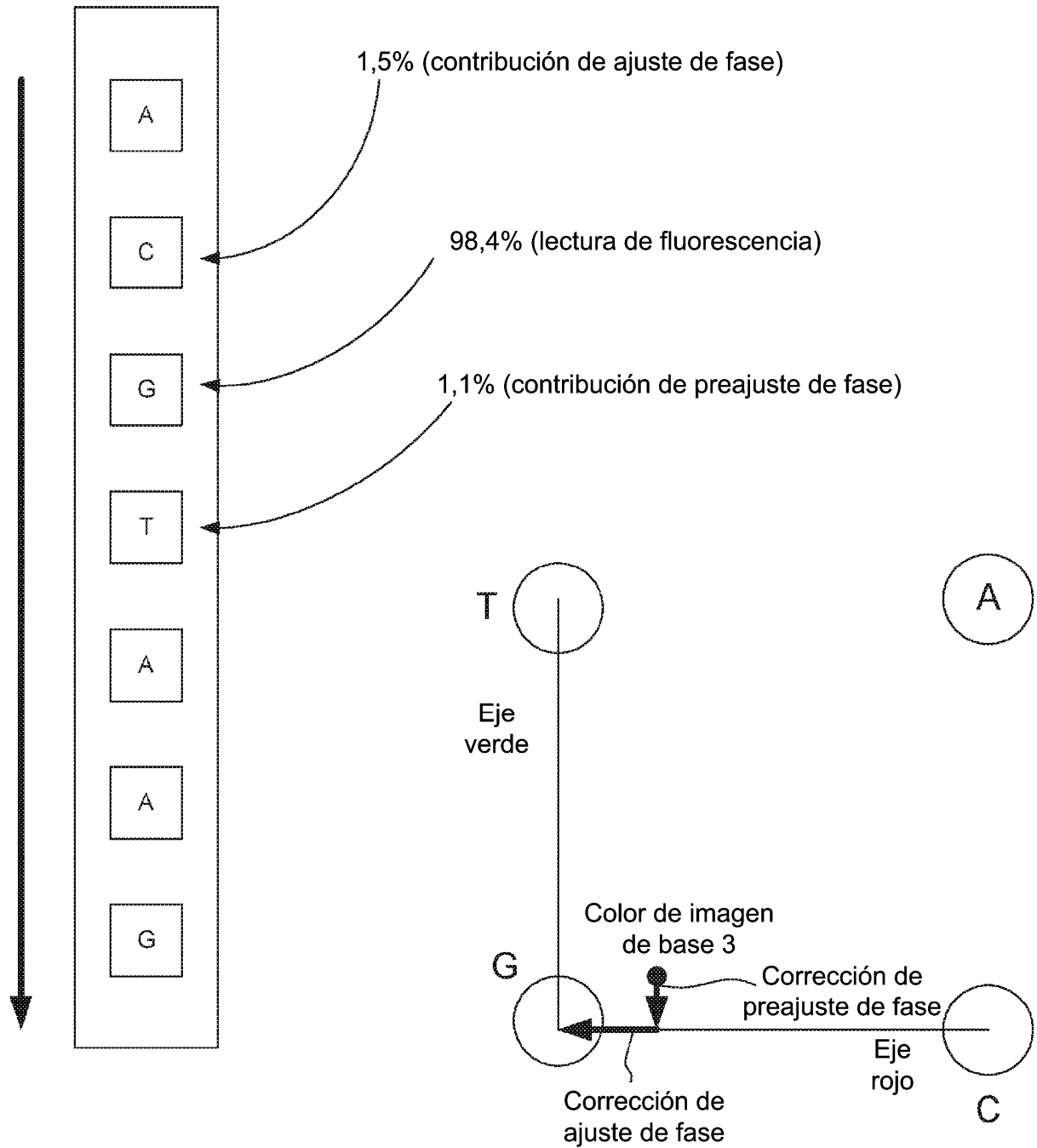
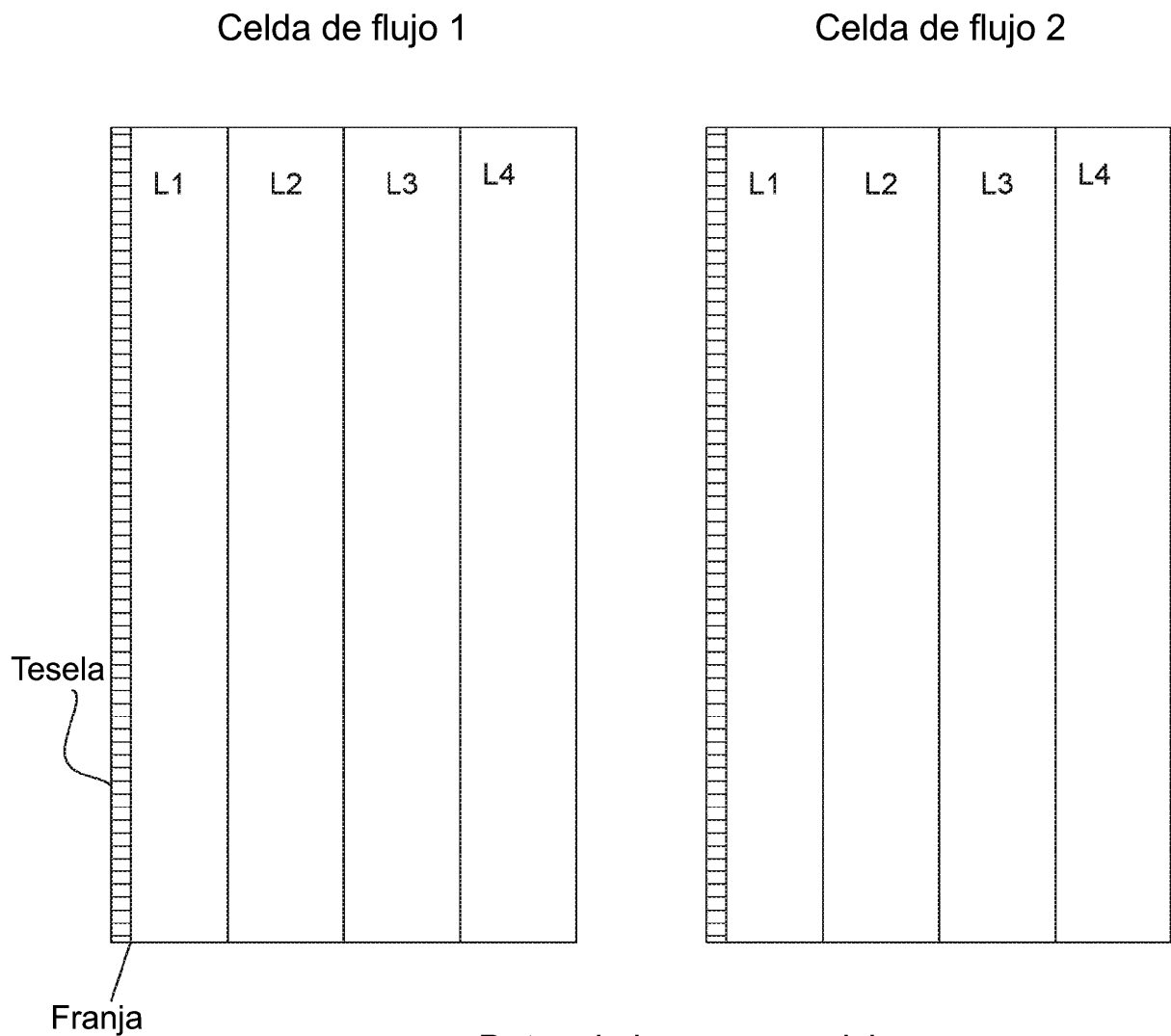


Figura 2



Datos de imagen por ciclo
procesados en tiempo real

2 Celdas de flujo
2 Superficies/celda de flujo
4 Carriles/Superficie
6 Franjas/Carril
120 Teselas/Franja

8000+ Teselas de datos por ciclo

Figura 3

Número de agrupaciones	1	2	3	4	5	6	7	8	9	10	11	12
Valor de Color (canal 1)	468	695	1255	521	753	1452	1355	1255	598	720	1193	422
Valor de Color (canal 2)	1321	1466	1326	498	463	1384	632	558	452	1467	1369	1278

Figura 4

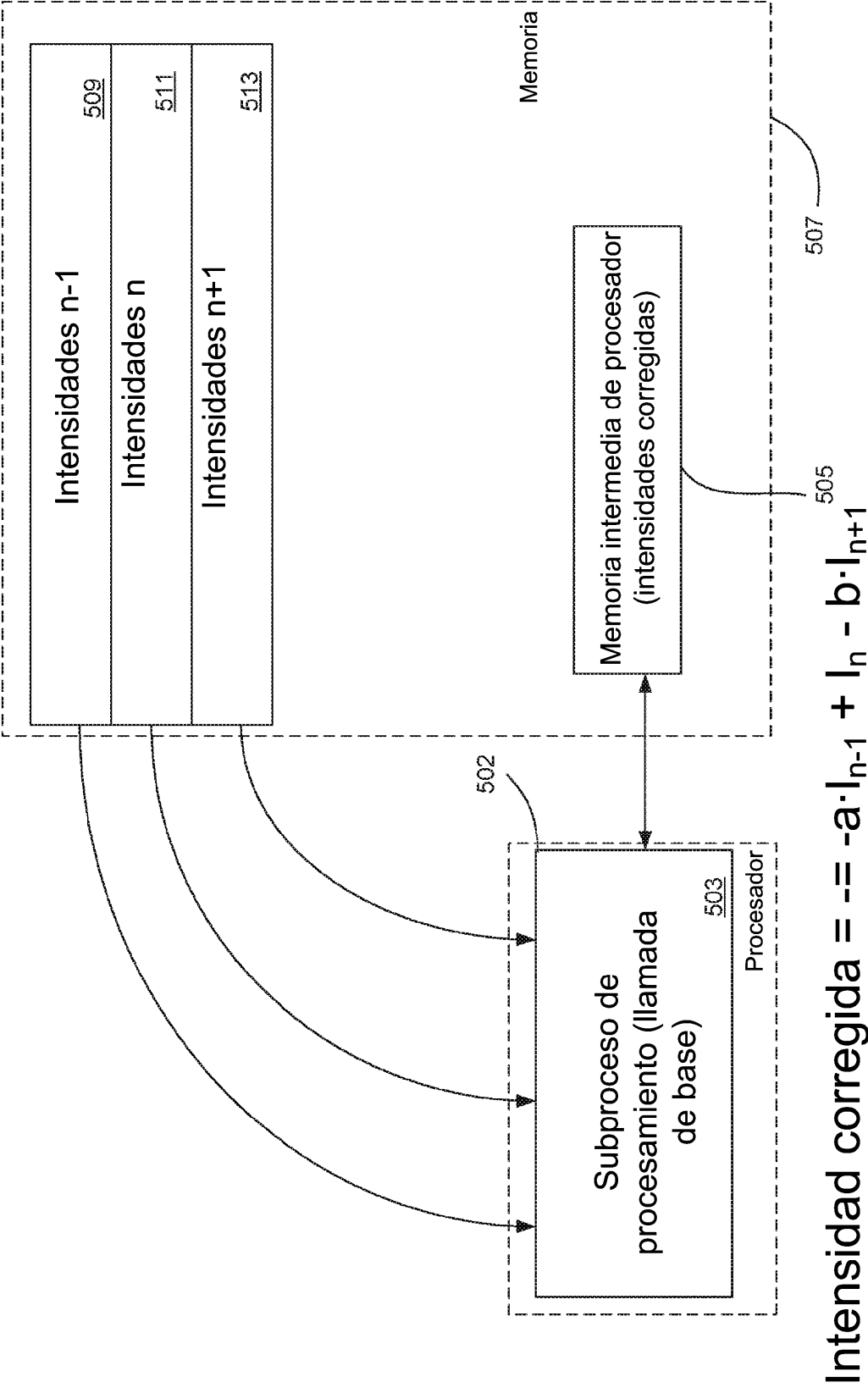


Figura 5

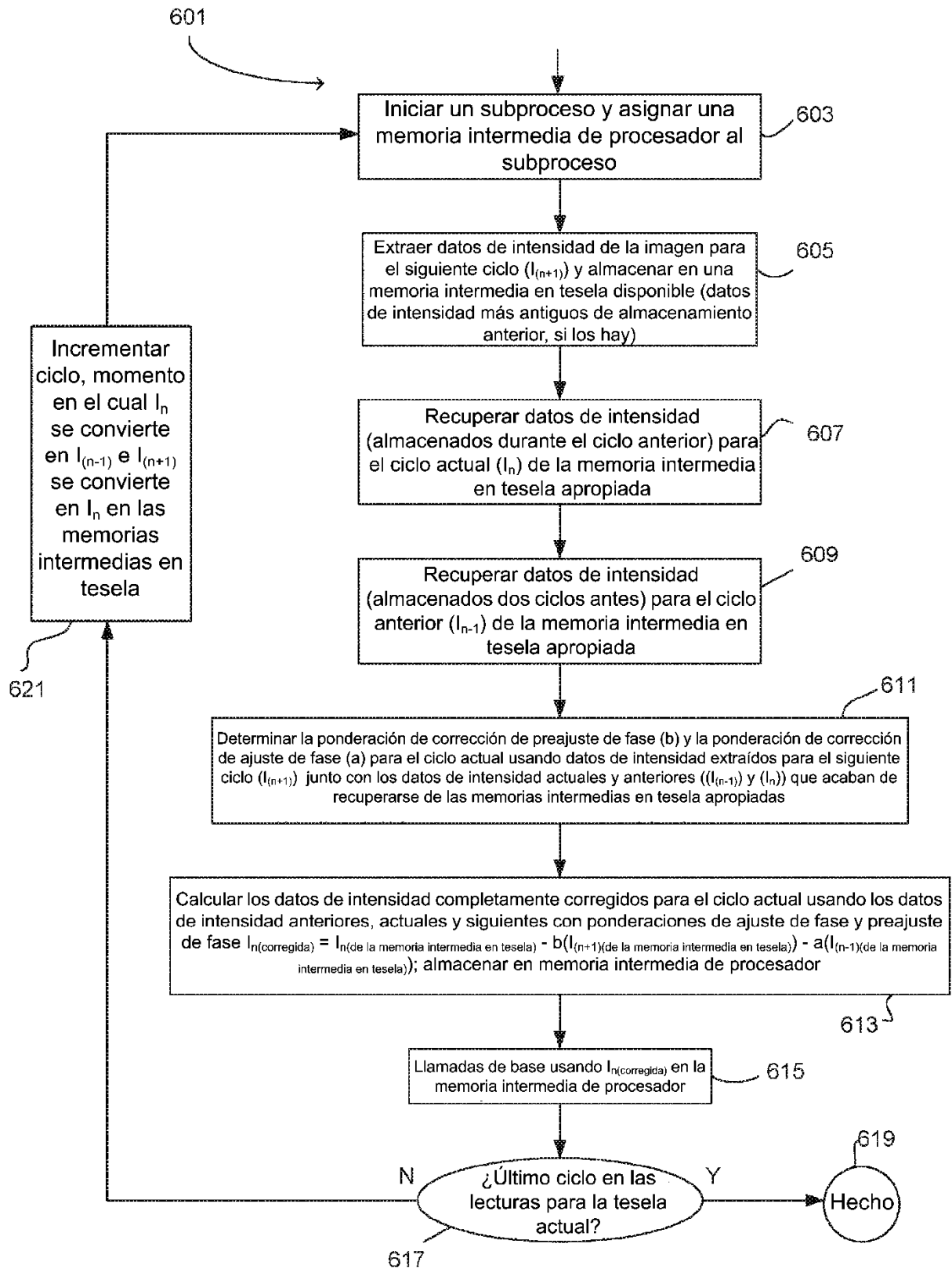


Figura 6

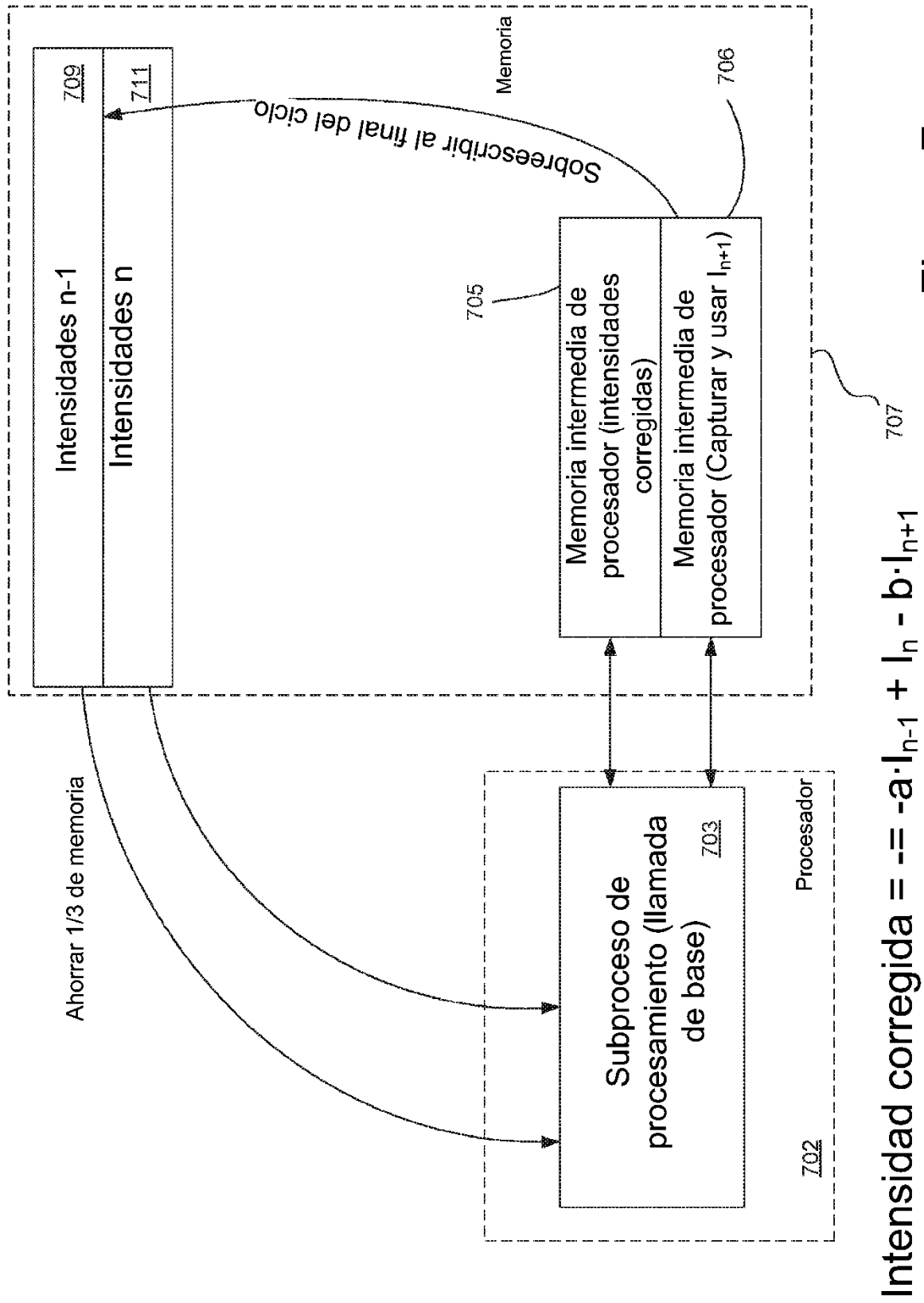
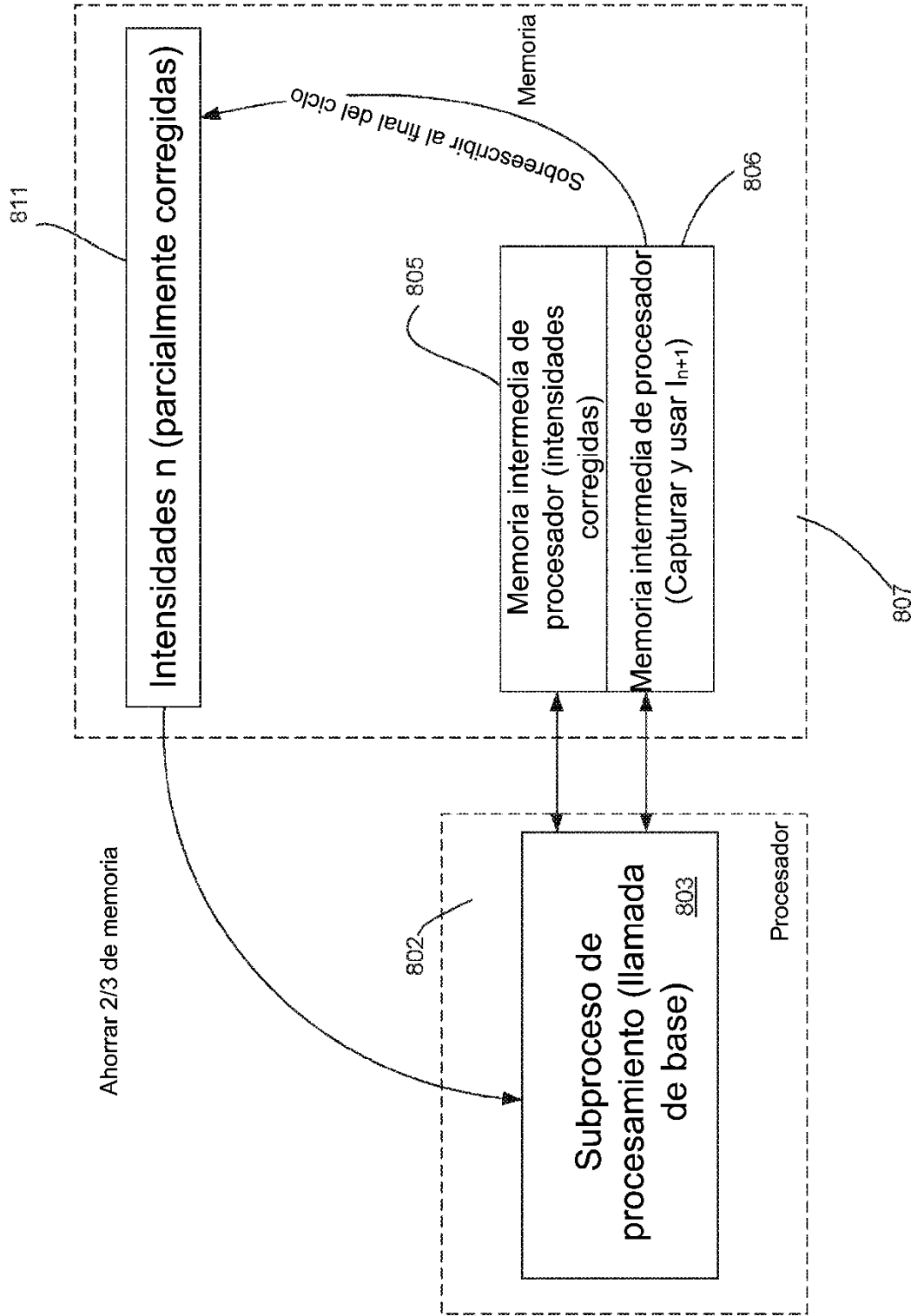


Figura 7



$$\text{Intensidad corregida} = -a \cdot I_{n-1} + I_n - b \cdot I_{n+1}$$

Figura 8

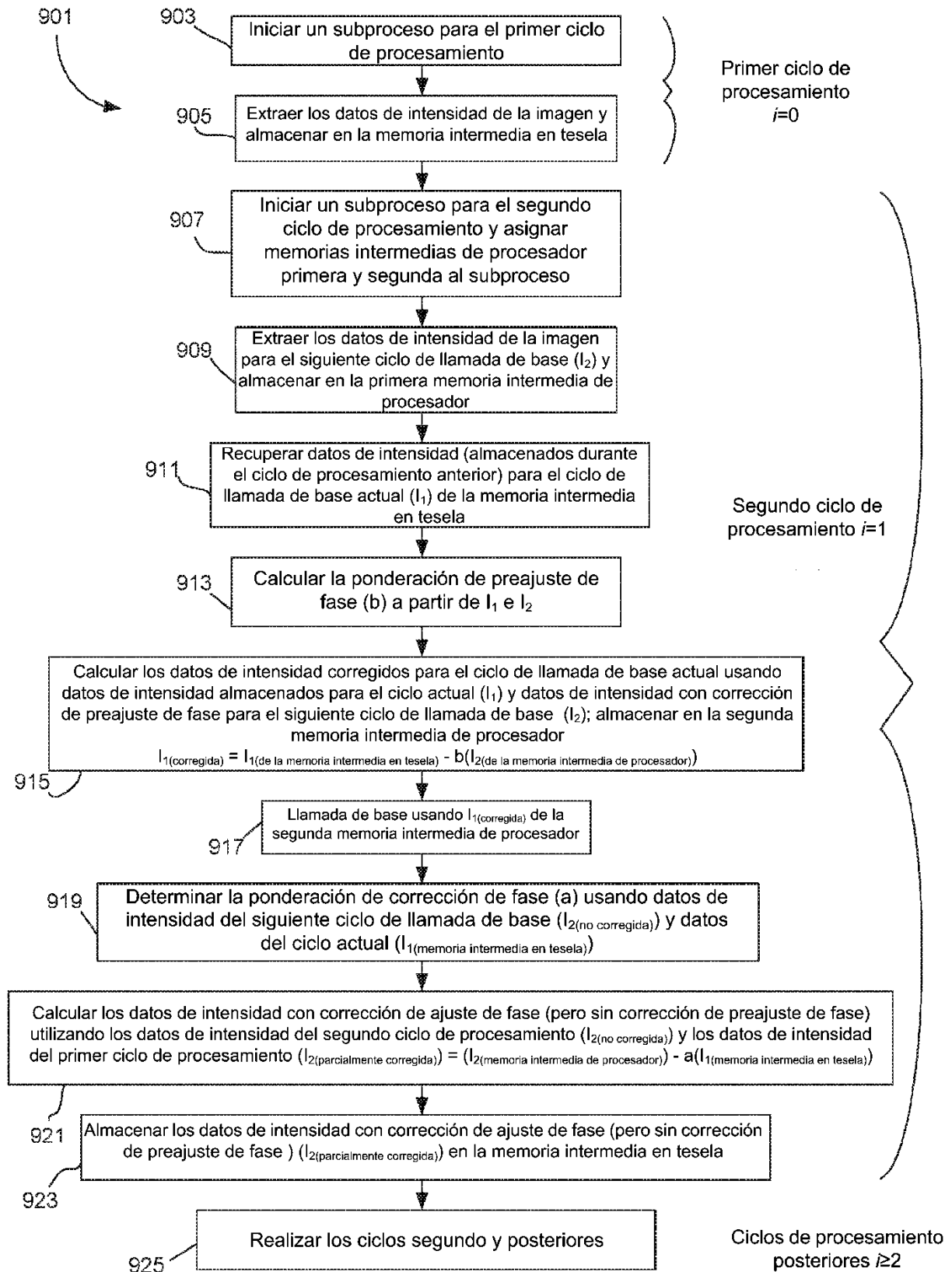


Figura 9

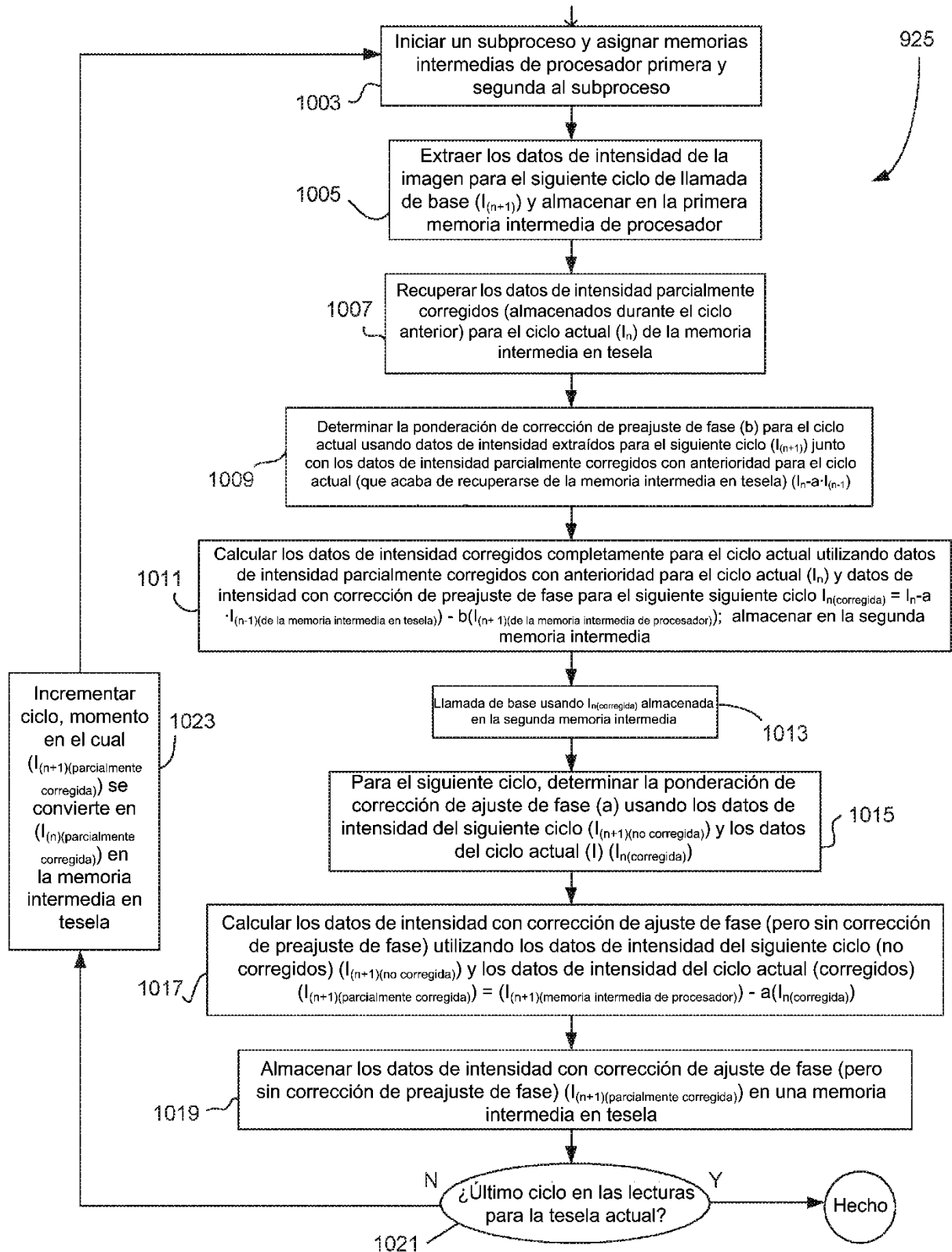


Figura 10

	Densidad (K/mm ²)	Agrupaciones PF (%)	Tasa de error de los ciclos	Alineado (%)	Tasa de error (%)
Referencia	206 +/- 0	81,32 +/- 0,69	150	1,43 +/- 0,01	1,38 +/- 0,00
Nuevo	206 +/- 0	81,80 +/- 0,00	150	1,42 +/- 0,00	1,41 +/- 0,00

	Teselas	Densidad (K/mm ²)	Agrupaciones PF (%)	Alineado (%)	Tasa de error (%)
Referencia	1	3815 +/- 0	65,39 +/- 0,00	1,43 +/- 0,00	1,89 +/- 0,00
Nuevo	1	3815 +/- 0	65,39 +/- 0,00	1,43 +/- 0,00	1,95 +/- 0,00

Figura 11