(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 704 255**

(51) Int. CI.:

C12Q 1/6869 (2008.01)
G06F 19/22 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **13.03.2013 PCT/US2013/030867**

(87) Fecha y número de publicación internacional: **18.09.2014 WO14142831**

(96) Fecha de presentación y número de la solicitud europea: **13.03.2013 E 13712642 (1)**

(97) Fecha y número de publicación de la concesión europea: **17.10.2018 EP 2971069**

(54) Título: **Métodos y sistemas para alinear elementos de ADN repetitivos**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
15.03.2019

(73) Titular/es:

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

(72) Inventor/es:

BRUAND, JOCELYNE;
RICHARDSON, TOM y
MANN, TOBIAS

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Métodos y sistemas para alinear elementos de ADN repetitivos

Antecedentes

Los conjuntos de elementos de ADN repetitivos polimórficos son útiles para muchas aplicaciones genéticas, incluyendo las pruebas de paternidad, la identificación humana (análisis forense de ADN), el control de quimeras (control de trasplantes de tejidos), así como muchos otros usos en la genómica de plantas y animales. Una clase de estos elementos repetitivos comprende las repeticiones cortas en tándem (STR, por sus siglas en inglés). El alelo de un locus STR se define por su longitud, o número de unidades repetidas, y por su variación de secuencia. Si bien los sistemas de electroforesis capilar pueden mostrar la longitud del alelo, las tecnologías de secuenciación tienen la capacidad adicional de diferenciación para descubrir la variación de la secuencia, tal como los SNP.

Para aprovechar los datos de NGS, es ventajoso asignar las lecturas de manera precisa y eficiente al locus STR y al alelo STR correctos. En Gymrek M. et al, 2012, *Genome Research*, 22: 1154-1162, se describe un método para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueadora conservada y una segunda región flanqueadora conservada.

Los métodos existentes para alinear las lecturas de secuenciación llevan tiempo y no son capaces de detectar todas las regiones repetitivas polimórficas conocidas y sin descubrir. Por tanto, existe una gran necesidad de métodos y sistemas mejorados para alinear elementos de ADN repetitivos.

Breve compendio

En esta memoria se presentan métodos y sistemas para alinear elementos de ADN repetitivos. Los métodos y sistemas utilizan los flancos conservados de loci polimórficos repetitivos para determinar efectivamente la longitud y la secuencia del elemento de ADN repetitivo.

Por consiguiente, una realización que se presenta en esta memoria es un método para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada, comprendiendo dicho método: (a) proporcionar un conjunto de datos que comprende al menos una lectura de secuencia del elemento de ADN repetitivo polimórfico; (b) proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada; (c) alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia; (d) alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y (e) determinar la longitud y/o secuencia de la región de repetición; en donde al menos las etapas (c), (d) y (e) se realizan utilizando un programa informático adecuado; y en donde el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende: (i) determinar una ubicación de una región flanqueante conservada en la lectura utilizando una coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y (ii) alinear la región flanqueante con la lectura de secuencia; en donde la región de siembra comprende una región de alta complejidad de la región flanqueante conservada, y la región de alta complejidad comprende una secuencia que es suficientemente distinta de la región de repetición para evitar el desalineamiento. En algunas realizaciones, el alineamiento puede comprender además alinear tanto la secuencia flanqueante como una región adyacente corta que comprende una porción de la región de repetición.

En esta memoria también se presenta un sistema para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada, comprendiendo dicho sistema: un procesador; y un programa para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico, comprendiendo el programa las instrucciones para que el procesador lleve a cabo las siguientes etapas: (a) proporcionar un conjunto de datos que comprende al menos una lectura de secuencia del elemento de ADN repetitivo polimórfico; (b) proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada; (c) alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia; (d) alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y (e) determinar la longitud y/o secuencia de la región de repetición; en donde el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende: (i) determinar una ubicación de una región flanqueante conservada en la lectura utilizando una coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y (ii) alinear la región flanqueante con la lectura de secuencia; en donde la región de siembra comprende una región de alta complejidad de la región flanqueante conservada, y la región de alta complejidad comprende una secuencia que es lo suficientemente distinta de la región de repetición para evitar el desalineamiento. En algunas realizaciones, el alineamiento puede comprender además alinear tanto la secuencia flanqueante como una región corta adyacente que comprende una porción de la región de repetición.

En ciertas realizaciones de los métodos o sistemas anteriores, la región de siembra comprende una región de alta complejidad de la región flanqueante conservada, por ejemplo, la región de alta complejidad que comprende una

secuencia que tiene una mezcla diversa de bases. En algunas realizaciones, la región de siembra evita las regiones de baja complejidad de la región flanqueante conservada, por ejemplo, una secuencia que sustancialmente se asemeja a la de la secuencia de repetición y/o la secuencia que tiene una mezcla de bases con baja diversidad.

En ciertas realizaciones de los métodos o sistemas anteriores, la región de siembra es directamente adyacente a la región de repetición y/o comprende una porción de la región de repetición. En ciertas realizaciones, la región de siembra está desplazada desde la región de repetición.

En ciertas realizaciones de los métodos o sistemas anteriores, el conjunto de datos de lecturas de secuencia comprende datos de secuencia a partir de un amplicón de PCR que tiene una secuencia de cebador directo e inverso. En ciertas realizaciones, al menos una lectura de secuencia en el conjunto de datos comprende una secuencia de consenso derivada de múltiples lecturas de secuencia. En ciertas realizaciones, proporcionar una secuencia de referencia comprende identificar un locus de interés en base a la secuencia de cebador del amplicón de PCR.

En ciertos métodos o sistemas, la región de repetición es una repetición corta en tándem (STR) tal como, por ejemplo, una STR seleccionada de los loci STR autosómicos de CODIS, loci Y-STR de CODIS, loci STR autosómicos EU, loci Y-STR EU y similares.

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

Breve descripción de los dibujos

La Figura 1 es un esquema que muestra un método de alineamiento de acuerdo con una realización.

La Figura 2 es un esquema que muestra varios errores de desalineamiento que pueden ocurrir si se utiliza la región flanqueante inmediatamente adyacente a la STR para sembrar el alineamiento.

La Figura 3 es un conjunto de gráficos que muestran la asignación de STR real en comparación con resultados teóricos basados en la entrada de muestras a partir de una mezcla de muestras.

La Figura 4 es una tabla que muestra 100% de coincidencia para las asignaciones alélicas de loci conocidos de cinco muestras de ADN de control.

Descripción detallada

Los conjuntos de elementos de ADN polimórficos y repetitivos son útiles para muchas aplicaciones genéticas, incluidas las pruebas de paternidad, la identificación humana (análisis forense de ADN), el control de quimeras (control de trasplantes de tejidos), así como muchos otros usos en la genómica de plantas y animales. Con el fin de aprovechar los datos de secuenciación de la próxima generación (NGS), se necesitan herramientas para la asignación precisa y eficiente de lecturas de secuenciación a los correctos locus y alelo de los elementos de ADN repetitivos. Una clase de estos elementos repetitivos comprende las repeticiones cortas en tándem (STR). El alelo de un locus STR se define por su longitud, o número de unidades repetidas, y por su variación de secuencia. Si bien los sistemas de electroforesis capilar pueden mostrar la longitud del alelo, las tecnologías de secuenciación tienen la capacidad adicional de diferenciación para descubrir la variación de las secuencias, tal como las SNP. Se apreciará que aunque los métodos y sistemas descritos aquí se consideran en el contexto de las STR, los mismos pueden aplicarse a cualquier otro elemento de ADN repetitivo.

Los métodos de alineamiento existentes fallan por varias razones. Un enfoque común es que frecuentemente se realiza el alineamiento con una secuencia de referencia. Sin embargo, la diferencia en el tamaño de los alelos difiere enormemente, incluso dentro de un solo locus. Por ejemplo, un locus de núcleo de EE.UU., FGA, tiene alelos conocidos entre 12,2 y 51,2, que incluyen diferencias de 156 nucleótidos (o incluso más). La mayoría de los alineadores no alinearán las lecturas con un espacio tan grande, y cualquier alelo que esté demasiado lejos de una secuencia de referencia será descartado por el alineador.

Otro enfoque con inconvenientes es el método de alineamiento con una escalera de referencia. Normalmente, se crea un "genoma de referencia" al construir una escalera de todos los alelos de STR conocidos y alinear las lecturas con esta referencia, como se hace típicamente con los datos de la secuencia del genoma completo de NGS o la secuenciación dirigida de regiones de ADN no repetitivas. Es método tiene deficiencias. Por ejemplo, se ignora la información conocida sobre la secuencia de STR, tal como la secuencia del cebador o las regiones flanqueantes conservadas. Las escaleras existentes son incompletas, ya que las secuencias de muchas regiones repetitivas polimórficas son actualmente desconocidas. Debido a la naturaleza altamente variable de estas regiones genómicas, se pueden descubrir nuevos alelos en el futuro. Además, los cambios en la secuencia de un alelo en la referencia pueden tener efectos globales en el alineamiento de las lecturas debido a la homología entre las secuencias.

Otra metodología alternativa para detectar las STR, conocida como lobSTR, detecta a todas las STR existentes a partir de los datos de secuenciación de una sola muestra *de novo*, sin conocimiento previo de las STR (ver Gyrnrek et al. 2012 *Genome Research* 22: 1154-62). Sin embargo, el método lobSTR ignora el conocimiento previo (secuencias

de cebadores, regiones flanqueantes) y recurre erróneamente a algunos alelos. Además, el lobSTR pierde los loci STR con patrones de repetición complejos, incluidos algunos de CODIS como D21S11, alelo 24 ($[TCTA]_4[TCTG]_6[TCTA]_3TA\ [TCTA]_3TCA[TCTA]_2TCCA\ TA[TCTA]_6$) o vWA, alelo 16 ($TCTA\ [TCTG]_3\ [TCTA]_{12}TCCA\ TCTA$). Además, el lobSTR asume alelos homocigotos o heterocigotos, y por lo tanto no es útil para el manejo de muestras que tienen mezclas.

Por lo tanto, existe una gran necesidad de un enfoque dirigido que utilice el conocimiento previo para aumentar en gran medida la sensibilidad y la especificidad.

En el presente documento se presentan métodos y sistemas que utilizan los flancos conservados de loci polimórficos repetitivos para determinar efectivamente la secuencia del elemento de ADN repetitivo. Los métodos alinean ventajosamente el comienzo de la secuencia de lectura con las posibles secuencias de cebadores para establecer el locus y la cadena a la que corresponde la lectura. Luego, las secciones de las secuencias flanqueantes apropiadas en cada lado del locus repetitivo se alinean con la lectura para extraer la longitud y secuencia exactas de la lectura. Estos alineamientos se siembran utilizando una estrategia de k-meros. Las regiones de siembra pueden estar, por ejemplo, en una región de alta complejidad preseleccionada de la secuencia flanqueante, cerca de la región de repetición, pero evitando la secuencia de baja complejidad con homología con el locus objetivo. Este enfoque evita ventajosamente el desalineamiento de secuencias flanqueantes de baja complejidad cerca de la región de repetición de interés.

El enfoque descrito en esta memoria es novedoso y sorprendentemente efectivo para determinar correctamente el tamaño y la secuencia del alelo. Los métodos emplean secuencias conocidas en los flancos de los propios STR, que se han definido previamente en base a las variaciones existentes conocidas entre la población humana. Ventajosamente, el alineamiento de un corto tramo de las regiones flanqueantes es computacionalmente rápido en comparación con otros métodos. Por ejemplo, un alineamiento de programación dinámica (tipo Smith-Waterman) de la lectura completa requiere el uso intensivo de CPU, y consume mucho tiempo, especialmente cuando se deben alinear múltiples lecturas de secuencia. Además, el tiempo dedicado a alinear una secuencia completa (para la que ni siquiera existe una referencia) requiere recursos informáticos valiosos.

El uso de regiones flanqueantes para determinar correctamente el alelo proporciona otras varias ventajas inesperadas con respecto a los métodos existentes. Por ejemplo, BWA, un alineador típico, funciona mal cuando se emplea para alinear con una referencia, principalmente debido a la naturaleza repetitiva de una secuencia STR y al estado incompleto de la referencia.

Además, los inventores han observado que cambiar la referencia para un locus STR a menudo afectó las asignaciones para otro locus, que debería ser independiente. Sin embargo, debido a que las aplicaciones forenses requieren asignaciones de alta fiabilidad, hay muy poco margen de error.

Realizaciones adicionales de los métodos proporcionados aquí identifican semillas únicas dentro de una secuencia flanqueante. Este enfoque permite reducir el tiempo de alineamiento y desempeña una función para evitar desalineamientos en el caso de flancos de baja complejidad.

Los métodos presentados aquí hacen uso del conocimiento previo de la secuencia flanqueante para asegurar la correcta asignación del alelo de STR. En contraste, los métodos existentes, que se basan en una secuencia de referencia completa para cada alelo, enfrentan tasas de fallo significativas en situaciones donde hay una referencia incompleta. Existen muchos alelos para los que no se conoce la secuencia, y posiblemente algunos alelos aún desconocidos. A modo de ilustración, se supone un locus con un patrón de repetición simple [TCTA] y un flanco 3' que se inicia con la secuencia TCAGCTA. Por lo tanto, la referencia puede incluir secuencias como [flanco1] $[TCTA]_n$TCAGCTA [resto_del_flanco2], en donde n es el número de repeticiones en el alelo. El alelo 9.3 diferiría del alelo 10 por tener una deleción en algún lugar de la secuencia. Con suerte, estos se incluirían en la referencia, aunque no todos podrían estarlo. $[TCTA]_7TCA\ [TCTA]_2$ es un ejemplo de tal alelo. Conforme a los protocolos de alineamiento existentes, cualquier lectura que finalice después del $[TCTA]_7$ y antes del [TCTA] final, se alineará con [flanco1] $[TCTA]_7$TCAGCTA, por lo que se realizará una asignación incorrecta.

Métodos de alineamiento

Los métodos proporcionados aquí permiten determinar la longitud de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada. Un método comprende proporcionar un conjunto de datos que comprende al menos una lectura de secuencia de un elemento de ADN repetitivo polimórfico; proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada; alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia; alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y determinar la longitud y/o secuencia de la región de repetición. En los métodos típicos, uno o más etapas en el método se realizan utilizando un programa informático adecuado.

Como se utiliza en esta memoria, el término "lectura de secuencia" se refiere a los datos de secuencia para los cuales se determinará la longitud y/o la identidad del elemento repetitivo. La lectura de secuencia puede comprender todo el

elemento repetitivo, o una porción del mismo. La lectura de secuencia puede comprender además una región flanqueante conservada en un extremo del elemento repetitivo (por ej., una región flanqueante 5'). La lectura de secuencia puede comprender además una región flanqueante adicional conservada en otro extremo del elemento repetitivo (por ej., una región flanqueante de 3'). En las realizaciones típicas, la lectura de secuencia comprende datos de secuencia de un amplicón de PCR que tiene una secuencia de cebador directo e inverso. Los datos de secuencia se pueden obtener a partir de cualquier metodología de secuencia adecuada. La lectura de la secuenciación puede ser, por ejemplo, a partir de una reacción de secuenciación por síntesis (SBS), una reacción de secuenciación por ligación, o cualquier otra metodología de secuenciación adecuada por la cual se desea determinar la longitud y/o la identidad de un elemento repetitivo. La lectura de secuencia puede ser una secuencia de consenso derivada de múltiples lecturas de secuencia. En ciertas realizaciones, proporcionar una secuencia de referencia comprende identificar un locus de interés en base a la secuencia del cebador del amplicón de PCR.

Como se utiliza en esta memoria, el término "elemento de ADN repetitivo polimórfico" se refiere a cualquier secuencia de ADN que se repite, y los métodos aquí proporcionados se pueden emplear para alinear las regiones flanqueantes correspondientes de cualquier secuencia de ADN que se repite. Los métodos presentados aquí se pueden utilizar para cualquier región de repetición. Los métodos presentados en la presente memoria se pueden usar para cualquier región que sea difícil de alinear, independientemente de la clase de repetición. El método presentado en esta memoria es especialmente útil para una región que tiene regiones flanqueantes conservadas. Adicionalmente o en forma alternativa, los métodos presentados en esta memoria son especialmente útiles para lecturas de secuenciación que abarcan toda la región de repetición, incluyendo al menos una porción de cada región flanqueante. En realizaciones típicas, el ADN repetitivo es una repetición en tándem de número variable (VNTR, por sus siglas en inglés). Las VNTR son polimorfismos en los que una secuencia particular se repite en ese locus muchas veces. Algunas VNTR incluyen minisatélites y microsatélites, también conocidos como repeticiones de secuencia simple (SSR, por sus siglas en inglés) o repeticiones cortas en tándem (STR, por sus siglas en inglés). En algunas realizaciones, la secuencia repetitiva es típicamente menor que 20 pares de bases, aunque se pueden alinear unidades que se repiten más grandes. Por ejemplo, en realizaciones típicas, la unidad que se repite puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 o más nucleótidos, y se puede repetir hasta 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 o hasta por lo menos 100 veces o más. En ciertas realizaciones, el elemento de ADN repetitivo polimórfico es una STR. En algunos métodos, la STR se utiliza con fines forenses. En métodos típicos para aplicaciones forenses, por ejemplo, el elemento de ADN repetitivo polimórfico comprende unidades de repetición de tetra o pentanucleótidos, sin embargo, los métodos proporcionados en esta memoria son adecuados para cualquier longitud de unidad de repetición. En ciertos métodos, la región de repetición es una repetición corta en tándem (STR) tal como, por ejemplo, una STR seleccionada de los loci STR autosómicos de CODIS, loci Y-STR de CODIS, loci STR autosómicos EU, loci Y-STR EU y similares. Como un ejemplo, la base de datos CODIS (Sistema de índice combinado de ADN) es un conjunto de loci STR centrales identificados por el laboratorio del FBI e incluye 13 loci: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 y D21S11. Las STR adicionales de interés para la comunidad forense y que se pueden alinear utilizando los métodos y sistemas proporcionados en esta memoria incluyen PENTA D y PENTA E. Los métodos y sistemas presentados aquí pueden aplicarse a cualquier elemento de ADN repetitivo y no se limitan a las STR descritas anteriormente. Como se emplea en esta memoria, el término "secuencia de referencia" se refiere a una secuencia conocida que actúa como un armazón sobre el cual se puede alinear una secuencia de muestra. En realizaciones típicas de los métodos y sistemas proporcionados en esta memoria, la secuencia de referencia comprende al menos una primera región flanqueante conservada y una segunda región flanqueante conservada. El término "región flanqueante conservada" se refiere a una región de secuencia fuera de la región de repetición. La región se conserva típicamente entre muchos alelos, aunque la región de repetición puede ser polimórfica. Una región flanqueante conservada como se usa en esta memoria típicamente será de mayor complejidad que la región de repetición. En realizaciones típicas, se puede emplear una única secuencia de referencia para alinear todos los alelos dentro de un locus. En algunas realizaciones, se emplea más de una secuencia de referencia para alinear todos los alelos dentro de un locus debido a la variación dentro de la región flanqueante. Por ejemplo, la región de repetición para Amelogenina tiene diferencias en los flancos entre X e Y, aunque una sola referencia puede representar la región de repetición si se incluye una región más larga en la referencia.

En los métodos presentados en esta memoria, una porción de una región flanqueante de una secuencia de referencia se alinea con la lectura de secuencia. El alineamiento se realiza determinando una ubicación de la región flanqueante conservada y luego realizando un alineamiento de secuencia de esa porción de la región flanqueante con la porción correspondiente de la lectura de secuencia. El alineamiento de una porción de una región flanqueante se realiza de acuerdo con métodos de alineamiento conocidos. En ciertos métodos, el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende: (i) determinar una ubicación de una región flanqueante conservada en la lectura mediante el uso de coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y (ii) alinear la región flanqueante con la lectura de secuencia. En algunos métodos, el alineamiento puede comprender además alinear tanto la secuencia flanqueante como una región corta adyacente que comprende una porción de la región de repetición.

Un ejemplo de este enfoque se ilustra en la Figura 1. En la Figura 1 se muestra un amplicón ("molde") que tiene una STR de longitud y/o identidad desconocida. Como se muestra en la Figura 1, se realiza un alineamiento inicial del cebador para identificar el locus de interés, en este caso una STR. Los cebadores se ilustran como p1 y p2, que son

las secuencias de cebadores que se utilizaron para generar el amplicón. En la realización que se muestra en la Figura 1, se usa solamente p1 durante la etapa de alineamiento del cebador. En algunas realizaciones, solamente p2 se utiliza para el alineamiento del cebador. En otras realizaciones, tanto p1 como p2 se usan para el alineamiento del cebador.

Después del alineamiento del cebador, se alinea el flanco 1, indicado como fl_{a1} en la Figura 1. El alineamiento del flanco 1 puede ir precedido por la siembra del flanco 1, designada como $f1_{siembra}$ en la Figura 1. Se siembra el flanco 1 para corregir un pequeño número (e) de indeles (inserciones o deleciones) entre el inicio de la lectura y la STR. La región de siembra puede estar directamente a continuación del inicio de la STR, o puede estar desplazada (como en la figura) para evitar regiones de baja complejidad. La siembra se puede realizar mediante coincidencia exacta de k-meros.

El alineamiento del flanco1 se continúa para determinar la posición inicial de la secuencia de STR. Si el patrón de STR se conserva lo suficiente como para predecir los primeros pocos nucleótidos (s1), estos se agregan al alineamiento para mejorar la precisión.

Como la longitud de la STR es desconocida, se realiza un alineamiento para el flanco 2 de la siguiente manera. Se realiza la siembra del flanco 2 para descubrir rápidamente las posibles posiciones finales de la STR. Como la siembra para el flanco 1, la siembra puede desplazarse para evitar regiones de baja complejidad y desalineamientos. Se desecha cualquier semilla del flanco 2 que no se alinee. Una vez que el flanco 2 se alinea correctamente, se puede determinar la posición final (s2) de la STR, y se puede calcular la longitud de la STR.

La región de siembra puede ser directamente adyacente a la región de repetición y/o comprender una porción de la región de repetición. En algunos métodos, la ubicación de la región de siembra dependerá de la complejidad de la región directamente adyacente a la región de repetición. El comienzo o el final de una STR puede estar limitado por una secuencia que comprende repeticiones adicionales o que tiene poca complejidad. Por lo tanto, puede ser ventajoso desplazar la siembra de la región flanqueante para evitar regiones de baja complejidad. Como se emplea en esta memoria, el término "baja complejidad" se refiere a una región con una secuencia que se asemeja a la de la secuencia de repetición. Adicionalmente o en forma alternativa, una región de baja complejidad incorpora una baja diversidad de nucleótidos. Por ejemplo, en algunas realizaciones, una región de baja complejidad comprende una secuencia que tiene más del 30%, 40%, 50%, 60%, 70% o más del 80% de similitud de secuencia con la secuencia de repetición. En realizaciones típicas, la región de baja complejidad incorpora cada uno de los cuatro nucleótidos a una frecuencia de menos del 20%, 15%, 10% o menos del 5% de todos los nucleótidos en la región. Se puede utilizar cualquier método adecuado para determinar una región de baja complejidad. Los métodos para determinar una región de baja complejidad son conocidos en la técnica, como se ejemplifica mediante los métodos descritos por Morgulis et al., (2006) *Bioinformatics*. 22 (2): 134-41. Por ejemplo, como se describe en Morgulis et al., se puede usar un algoritmo tal como DUST para identificar regiones dentro de una secuencia de nucleótidos dada que tienen baja complejidad.

En algunas realizaciones, la siembra se desplaza del inicio de la STR en al menos 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40 o más nucleótidos. En algunas realizaciones, la región flanqueante se evalúa para identificar una región de alta complejidad. Tal como se utiliza aquí, el término "región de alta complejidad" se refiere a una región con una secuencia que es suficientemente diferente de la de la repetición para eliminar las posibilidades de desalineamientos. Adicionalmente o en forma alternativa, una región de alta complejidad incorpora una variedad de nucleótidos. Por ejemplo, en algunas realizaciones, una región de alta complejidad comprende una secuencia que tiene menos del 80%, 70%, 60%, 50%, 40%, 30%, 20% o menos del 10% de similitud con la secuencia de repetición. En realizaciones típicas, la región de alta complejidad incorpora cada uno de los cuatro nucleótidos a una frecuencia de al menos 10%, 15%, 20% o al menos el 25% de todos los nucleótidos en la región.

Tal como se emplea en la presente memoria, el término "coincidencia exacta de k-meros" se refiere a un método para encontrar un alineamiento óptimo utilizando un método de palabra donde la longitud de la palabra se define por tener un valor *k*. En algunas realizaciones, el valor de *k* es 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23. , 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 o más nucleótidos de longitud. En las realizaciones típicas, k tiene un valor de entre 5 y 30 nucleótidos de longitud. En algunas realizaciones típicas, k tiene un valor de entre 5 y 16 nucleótidos de longitud. En ciertas realizaciones, k se selecciona en línea. Por ejemplo, si una región flanqueante es corta (cebador cerca de STR), k se reduce adecuadamente. En realizaciones típicas, k se selecciona para garantizar la búsqueda de todas las coincidencias con la distancia de edición e. Los métodos de palabras identifican una serie de subsecuencias cortas y no superpuestas ("palabras") en la secuencia de consulta que luego se comparan con las secuencias candidatas de la base de datos. Las posiciones relativas de la palabra en las dos secuencias que se comparan se eliminan para obtener un desplazamiento; esto indicará una región de alineamiento si múltiples palabras distintas producen el mismo desplazamiento. Solamente si se detecta esta región, estos métodos aplicarán criterios de alineamiento más sensibles, por lo que se eliminan muchas comparaciones innecesarias con secuencias sin una similitud apreciable. Los métodos para evaluar la coincidencia de k-meros, incluyendo la coincidencia exacta de k-meros, son bien conocidos en la técnica, como lo ejemplifican las divulgaciones de Lipman, et al., (1985) *Science* 227: 1435-41, y de Altschul, et al., (1990) *Journal of Molecular Biology* 215: 403-410.

En ciertas realizaciones, proporcionar una secuencia de referencia comprende identificar un locus de interés en base a la secuencia del cebador de un amplicón. Como se emplea en esta memoria, el término "amplicón" se refiere a

cualquier producto de amplificación adecuado para el cual se obtiene una secuencia. Típicamente, el producto de amplificación es un producto de una metodología de amplificación selectiva, que utiliza cebadores específicos de objetivos, tal como los cebadores de PCR. En ciertas realizaciones, los datos de secuencia son de un amplicón de PCR que tiene una secuencia de cebador directo e inverso. En algunas realizaciones, la amplificación selectiva puede incluir una o más etapas de amplificación no selectivas. Por ejemplo, un proceso de amplificación que utiliza cebadores aleatorios o degenerados puede ir seguido de uno o más ciclos de amplificación que emplean cebadores específicos de objetivo. Los métodos adecuados para la amplificación selectiva incluyen, pero no se limitan a, la reacción en cadena de la polimerasa (PCR, por sus siglas en inglés), la amplificación por desplazamiento de cadena (SDA, por sus siglas en inglés), la amplificación mediada por transcripción (TMA, por sus siglas en inglés) y la amplificación basada en la secuencia de ácido nucleico (NASBA, por sus siglas en inglés), tal como se describe en la patente de EE.UU. No. 8.003.354. Los métodos de amplificación anteriores pueden emplearse para amplificar selectivamente uno o más ácidos nucleicos de interés. Por ejemplo, la PCR, incluyendo PCR múltiplex, SDA, TMA, NASBA y similares, puede utilizarse para amplificar selectivamente uno o más ácidos nucleicos de interés. En tales realizaciones, los cebadores dirigidos específicamente al ácido nucleico de interés se incluyen en la reacción de amplificación. Otros métodos adecuados para la amplificación de ácidos nucleicos pueden incluir la extensión y la ligación de oligonucleótidos y la ligación, la amplificación de círculo rodante (RCA, por sus siglas en inglés) (Lizardi et al., Nat. Genet. 19: 225-232 (1998)) y el ensayo de ligación de oligonucleótidos (OLA, por sus siglas en inglés) (véanse en general las patentes de EE.UU. números 7.582.420, 5.185.243, 5.679.524 y 5.573.907; patente europea EP 0 320 308 B1; patente europea EP 0 336 731 B1; patente europea EP 0 439 182 B1; patente WO 90/01069; patente WO 89/12696; y patente WO 89/09835). Se apreciará que estas metodologías de amplificación pueden diseñarse para amplificar selectivamente un ácido nucleico objetivo de interés. Por ejemplo, en algunas realizaciones, el método de amplificación selectiva puede incluir reacciones de ensayo de amplificación de la sonda por ligación o por ligación de oligonucleótidos (OLA) que contienen cebadores dirigidos específicamente al ácido nucleico de interés. En algunas realizaciones, el método de amplificación selectiva puede incluir una reacción por ligación-extensión del cebador que contiene cebadores dirigidos específicamente al ácido nucleico de interés. Como ejemplo no limitativo de cebadores de extensión y cebadores de ligación que pueden diseñarse específicamente para amplificar un ácido nucleico de interés, la amplificación puede incluir cebadores utilizados para el ensayo GoldenGate™ (Illumina, Inc., San Diego, CA), como se describe en la patente de EE. UU. No. 7.582.420. Los presentes métodos no están limitados a ninguna técnica de amplificación particular y las técnicas de amplificación descritas en esta memoria son solo ejemplos con respecto a los métodos y realizaciones de la presente divulgación.

Los cebadores para la amplificación de un elemento de ADN repetitivo típicamente se hibridan con las secuencias únicas de las regiones flanqueantes. Los cebadores pueden diseñarse y generarse de acuerdo con cualquier metodología adecuada. El diseño de cebadores para regiones flanqueantes de regiones de repetición es bien conocido en la técnica, como se ejemplifica en Zhi, et al. (2006) *Genome Biol*, 7 (1): R7. Por ejemplo, los cebadores se pueden diseñar manualmente. Esto implica buscar repeticiones de microsatélite en la secuencia de ADN genómico, lo que se puede hacer mediante examen visual o con herramientas automatizadas como el software RepeatMasker. Una vez que se determinan las regiones de repetición y las regiones flanqueantes correspondientes, las secuencias flanqueantes se pueden usar para diseñar marcadores de oligonucleótidos que amplificarán la repetición específica en una reacción PCR.

Sistemas

En esta memoria también se presenta un sistema para determinar la longitud de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada, comprendiendo el sistema: un procesador; y un programa para determinar la longitud de un elemento de ADN repetitivo polimórfico, comprendiendo el programa instrucciones para: (a) proporcionar un conjunto de datos que comprenda al menos una lectura de secuencia del elemento de ADN repetitivo polimórfico; (b) proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada; (c) alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia; (d) alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y (c) determinar la longitud y/o secuencia de la región de repetición; en donde al menos las etapas (c), (d) y (e) se realizan utilizando un programa informático adecuado. En algunos sistemas, el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende: (i) determinar una ubicación de una región flanqueante conservada en la lectura utilizando la coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y (ii) alinear la región flanqueante con la lectura de secuencia. En algunos sistemas, el alineamiento puede comprender además alinear tanto la secuencia flanqueante como una región corta adyacente que comprende una porción de la región de repetición.

Un sistema capaz de llevar a cabo un método descrito en esta memoria puede, aunque no necesariamente, estar integrado con un dispositivo de secuenciación. Más bien, también es posible un sistema independiente o un sistema integrado con otros dispositivos. Un sistema capaz de llevar a cabo un método descrito en esta memoria, ya sea integrado o no con capacidades de detección, puede incluir un controlador del sistema que sea capaz de ejecutar un conjunto de instrucciones para llevar a cabo una o más etapas de un método, técnica o proceso descritos en esta memoria . Opcionalmente, las instrucciones pueden además dirigir la realización de las etapas para detectar ácidos nucleicos. Un controlador del sistema útil puede incluir cualquier sistema basado en procesador o basado en

microprocesador, incluidos los sistemas que utilizan microcontroladores, computadoras con un conjunto de instrucciones reducido (RISC, por sus siglas en inglés), circuitos integrados específicos de aplicación (ASIC, por sus siglas en inglés), matriz de compuerta programable en campo (FPGA, por sus siglas en inglés), circuitos lógicos y cualquier otro circuito o procesador capaz de ejecutar funciones descritas en esta memoria. Un conjunto de instrucciones para un controlador del sistema puede tener la forma de un programa de software. Como se emplean en esta memoria, los términos "software" y "firmware" son intercambiables, e incluyen cualquier programa de computación almacenado en la memoria para ser ejecutado por una computadora, incluyendo memoria RAM, memoria ROM, memoria EPROM, memoria EEPROM y memoria RAM no volátil (NVRAM, por sus siglas en inglés). El software puede estar en varias formas, tal como software del sistema o software de aplicación. Además, el software puede estar en forma de una colección de programas separados, o un módulo de programa dentro de un programa más grande o una porción de un módulo de programa. El software también puede incluir programación modular en forma de programación orientada a objetos.

Ejemplo 1

Alineamiento del locus D 18S51

Este ejemplo describe el alineamiento del locus D18S51 según una realización. Algunos loci tienen secuencias flanqueantes que son de baja complejidad y se asemejan a la secuencia de repetición STR. Esto puede hacer que la secuencia flanqueante no esté alineada (a veces con la propia secuencia de STR) y, por lo tanto, el alelo puede ser mal asignado. Un ejemplo de un locus problemático es D18S51. El motivo de repetición es **[AGAA]n AAAG AGAGAG**. La secuencia flanqueante se muestra a continuación con la secuencia de "problema" de baja complejidad subrayada:

GAGACCTTGTCTC **(STR)** **GAAAGAAAGAGAAAAAGAAAAGAAA**TAGTAGCAACTGTTAT

Si la región flanqueante inmediatamente adyacente a la STR se utilizara para sembrar el alineamiento, se generarían k-meros como GAAAG, AAAGAA, AGAGAAA, que mapean a la secuencia de STR. Esto impide el rendimiento, ya que se obtienen muchas posibilidades a partir de la siembra, pero lo más importante es que el enfoque crea desalineamientos, como los que se muestran en la Figura 2. En las secuencias que se muestran en la Figura 2, se resalta la correcta secuencia de STR, la secuencia de STR que se obtiene del desalineamiento está subrayada y los errores de lectura se muestran en negrita.

Para estos flancos de baja complejidad, se aseguró que las regiones de siembra no se encuentren en la región de baja complejidad al alejarlos de la secuencia de STR. Si bien esto requiere lecturas más largas para la asignación de la STR, se garantiza una alta precisión y se evita el desalineamiento de la región flanqueante con la secuencia de STR (u otras porciones del flanco). El flanco de baja complejidad aún está alineado con la lectura para encontrar la posición final de la STR, pero debido a que el alineamiento está sembrado con una secuencia de alta complejidad, el mismo tiene que estar en la posición correcta.

Ejemplo 2

Alineamiento del locus Penta-D mediante la adición de secuencia STR corta

Un conjunto de secuencias Penta-D tendió a tener STR que fueron 1 nt más cortos de lo esperado. Tras una inspección adicional, se descubrió que ambos flancos contenían segmentos de poli-A y que los errores de secuenciación/ amplificación a menudo eliminaban una de las A en esos segmentos. Como se muestra en la siguiente secuencia, los segmentos homopoliméricos A se encuentran en ambos flancos.

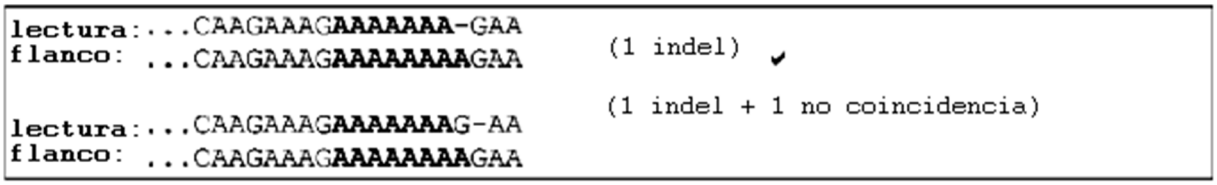
...CAAGAAAGAAAAAAAAG **[AAAGA] n** AAAAACGAAGGGGAAAAAAAGAGAAT ...

Un error de lectura que cause una deleción en el primer flanco produciría hasta dos alineamientos igualmente viables:

lectura:...CAAGAAAG**AAAAAAA**-GA...
flanco: ...CAAGAAAG**AAAAAAAA**G- (2 indeles)

lectura:...CAAGAAAG**AAAAAAA**GA... (2 no coincidencias)
flanco: ...CAAGAAAG**AAAAAAAA**G

Hacer que la base más cercana a la STR sea una coincidencia no funcionó porque uno de los flancos en uno de las STR terminó teniendo un SNP, lo que hizo que se reconsiderara ese método en su totalidad. Se descubrió que agregar solo 2 bases de la secuencia de STR resolvió la cuestión:



Ejemplo 3

Análisis de la mezcla de muestras de ADN

Se analizó una mezcla de muestras utilizando los métodos proporcionados en esta memoria para realizar asignaciones precisas para cada locus en un panel de STR forenses. Para cada locus, se contaron las lecturas de números correspondientes a cada alelo y a cada secuencia diferente para ese alelo.

Los resultados típicos se muestran en la Figura 3. Como se muestra, la barra a la derecha de cada par representa los datos reales obtenidos, que indican la proporción de lecturas para cada alelo. Los tonos diferentes representan secuencias diferentes. Se omiten los alelos con menos del 0,1% del recuento de lectura de locus y las secuencias con menos del 1% del recuento de alelos. La barra en el lado izquierdo de cada par representa las proporciones teóricas (sin intermitencias). Diferentes tonos representan diferentes ADN de control en la entrada como se indica en la leyenda. En la Figura 3, el eje x está en alelo de orden, y el eje Y indica la proporción de lecturas con el alelo indicado.

Como se muestra en la Figura, el enfoque de asignación de STR que emplea los métodos presentados en esta memoria lograron asignaciones sorprendentemente exactas para cada alelo en el panel.

Ejemplo 4

Análisis del panel de STR forense

Se analizó un panel de 15 loci diferentes en 5 muestras diferentes. Las muestras se obtuvieron de Promega Corp, e incluyeron las muestras 9947A, K562, 2800M, NIST: A y B (SRM 2391c). Los loci se eligieron entre los marcadores forenses STR de CODIS e incluyeron CSF1PO, D3S1358, D7S820, D16S539, D18S51, FGA, PentaE, TH01, vWA, D5S818, D8S1179, D13S317, D21S11, PentaD y TPOX empleando el método de alineamiento presentado en esta memoria. Brevemente, los marcadores se amplificaron utilizando cebadores estándar, como se describe en Krenke, et al. (2002) *J. Forensic Sci.* 47 (4): 773-785. Los amplicones se agruparon y los datos de secuenciación se obtuvieron utilizando ciclos de 1x460 en un instrumento de secuenciación MiSeq (Illumina, San Diego, CA).

El alineamiento se realizó de acuerdo con los métodos presentados en esta memoria. Como se indica en la Fig. 4, se mostró una coincidencia del 100% para estas muestras de control en comparación con los datos de control. Además, este método identificó un SNP previamente desconocido en una de las muestras para el marcador D8S1179, lo que además demuestra la poderosa herramienta del análisis de STR basado en la secuencia cuando se combina con los métodos de alineamiento aquí proporcionados.

Se pretende que el término “que comprende” en la presente memoria sea abierto, incluyendo no solo los elementos citados, sino que además abarque cualquier elemento adicional.

Se han descrito varias realizaciones. Sin embargo, se entenderá que pueden hacerse varias modificaciones.

REIVINDICACIONES

1. Un método para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada, comprendiendo dicho método:

(a) proporcionar un conjunto de datos que comprende al menos una lectura de secuencia del elemento de ADN repetitivo polimórfico;

(b) proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada;

(c) alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia;

(d) alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y

(e) determinar la longitud y/o secuencia de la región de repetición;

en donde al menos las etapas (c), (d) y (c) se realizan utilizando un programa informático adecuado;

en donde el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende:

(i) determinar una ubicación de una región flanqueante conservada en la lectura utilizando una coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y

(ii) alinear la región flanqueante con la lectura de secuencia; en donde la región de siembra comprende una región de alta complejidad de la región flanqueante conservada, comprendiendo la región de alta complejidad una secuencia que es suficientemente distinta de la región de repetición para evitar el desalineamiento.

2. El método de la reivindicación 1, que comprende además alinear tanto la secuencia flanqueante como una región corta adyacente que comprende una porción de la región de repetición.

3. El método de la reivindicación 1, en donde la región de alta complejidad comprende una secuencia que tiene una mezcla diversa de bases.

4. El método de la reivindicación 1, en donde la región de siembra evita las regiones de baja complejidad de la región flanqueante conservada.

5. El método de la reivindicación 4, en donde la región de baja complejidad comprende una secuencia que sustancialmente se asemeja a la de la secuencia de repetición.

6. El método de la reivindicación 4, en donde la región de baja complejidad comprende una secuencia que tiene una mezcla de bases con baja diversidad.

7. El método de la reivindicación 1, en donde la región de siembra es directamente adyacente a la región de repetición.

8. El método de la reivindicación 1, en donde la región de siembra comprende una porción de la región de repetición.

9. El método de la reivindicación 1, en donde la región de siembra está desplazada de la región de repetición.

10. El método de la reivindicación 1, en donde al menos una lectura de secuencia en el conjunto de datos comprende una secuencia de consenso derivada de múltiples lecturas de secuencia.

11. El método de la reivindicación 1, en donde proporcionar una secuencia de referencia comprende identificar un locus de interés en base a una secuencia de cebador de un amplicón de PCR.

12. Un sistema para determinar la longitud y/o secuencia de un elemento de ADN repetitivo polimórfico que tiene una región de repetición situada entre una primera región flanqueante conservada y una segunda región flanqueante conservada, comprendiendo el sistema:

un procesador; y

un programa para determinar la longitud y/o secuencia de un elemento polimórfico repetitivo de ADN, en donde el programa comprende instrucciones para que el procesador lleve a cabo las siguientes etapas:

(a) proporcionar un conjunto de datos que comprende al menos una lectura de secuencia del elemento polimórfico repetitivo de ADN;

(b) proporcionar una secuencia de referencia que comprende la primera región flanqueante conservada y la segunda región flanqueante conservada;

(c) alinear una porción de la primera región flanqueante de la secuencia de referencia con la lectura de secuencia;

(d) alinear una porción de la segunda región flanqueante de la secuencia de referencia con la lectura de secuencia; y

(e) determinar la longitud y/o secuencia de la región de repetición;

en donde el alineamiento de una porción de la región flanqueante en una o ambas etapas (c) y (d) comprende:

(i) determinar una ubicación de una región flanqueante conservada en la lectura utilizando una coincidencia exacta de k-meros de una región de siembra que se superpone o es adyacente a la región de repetición; y

(ii) alinear la región flanqueante con la lectura de secuencia;

en donde la región de siembra comprende una región de alta complejidad de la región flanqueante conservada, comprendiendo la región de alta complejidad una secuencia que es lo suficientemente distinta de la región de repetición para evitar el desalineamiento.

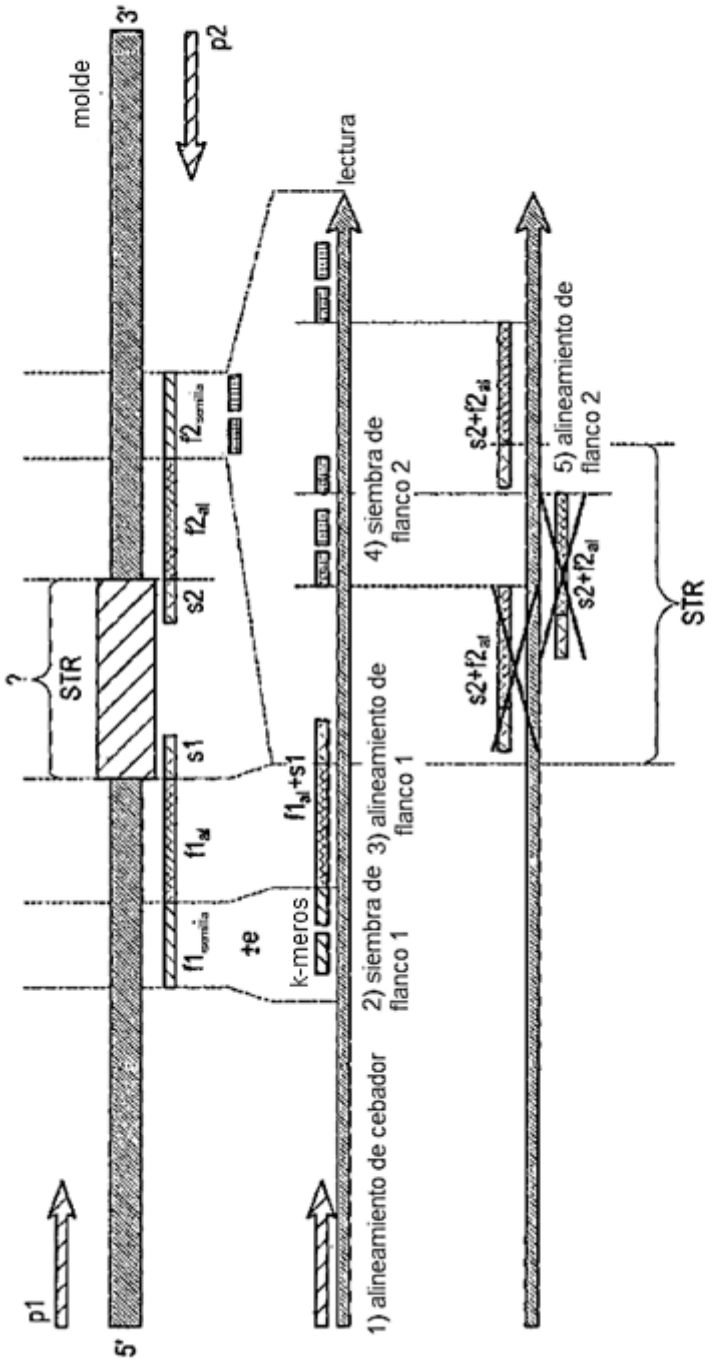

molde
3'
p2
lectura
STR
?
s2
s1
$f2_{semilla}$
$f2_{al}$
$f1_{semilla}$
$f1_{al}$
$f1_{al}$+s1
±e
k-meros
$s2+f2_{al}$
$s2+f2_{al}$
$s2+f2_{al}$
1) alineamiento de cebador
2) siembra de flanco 1
3) alineamiento de flanco 1
4) siembra de flanco 2
5) alineamiento de flanco 2
STR
p1
5'

Fig. 1

lectura :
...AGAAAGAAAGAAAGAAAGAAA
(debería ser demasiado corta para asignación)
flanco:
GAAAGAAAGAGAAAAAAGAAATAGTAGCAACTGTTAT...
lectura :
...AGAAAGAAAGAAAGAAAAAGAGAGGAAAGAAAGAGAAAAGAAAAGAAATAGTAG...
flanco 2:
GAAAGAAAGAGAAAAAAGAAAATAGTAGCAAC...
lectura :
...AGAAAGAAAAAGAGAGAGGAAAGAAAGAAAAAAAAAGAAAAAAAATAGTAG...
flanco 2:
GAAAGAAAGAGAAAAAAGAAAATA...

Fig. 2

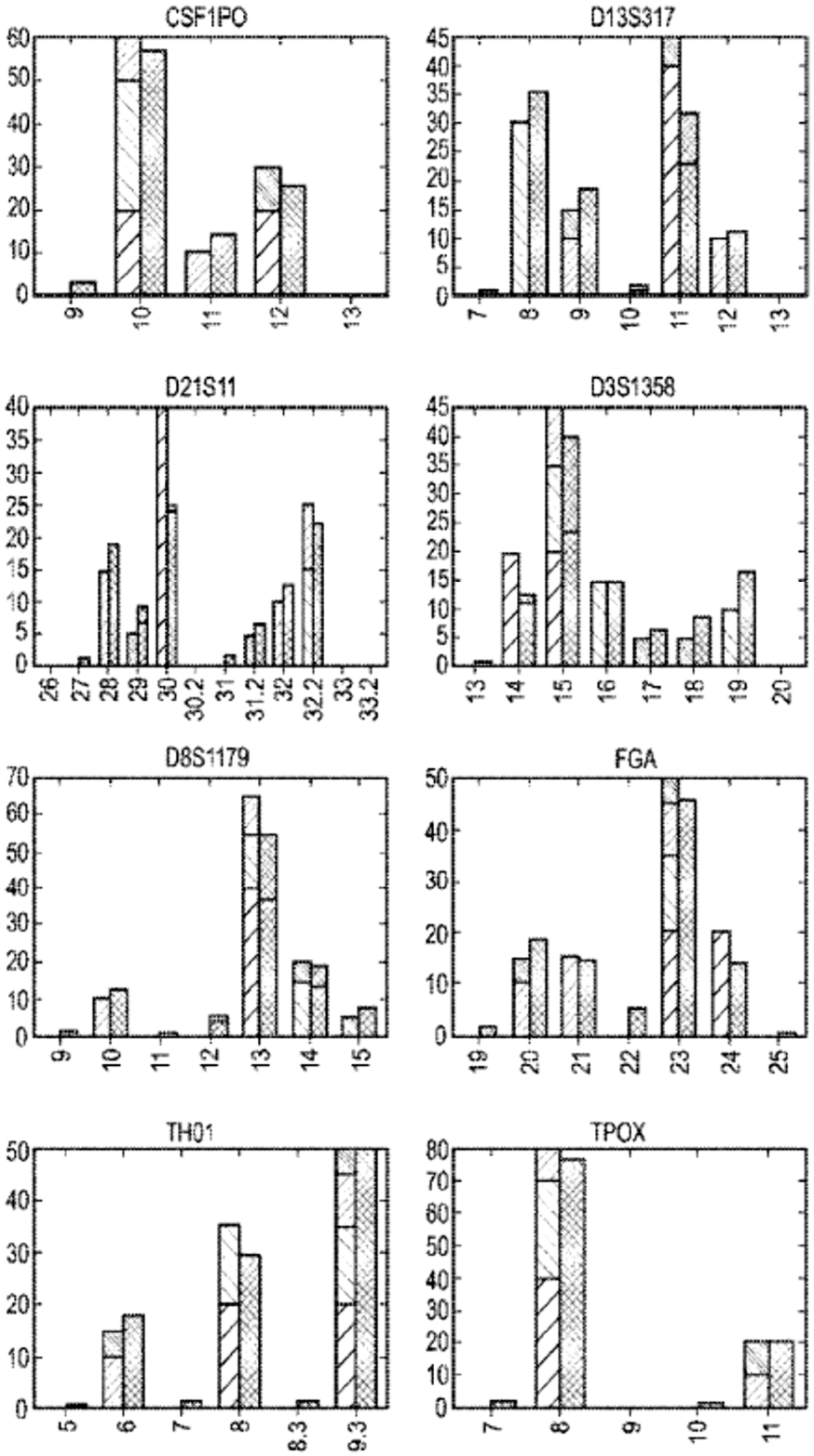


Fig. 3

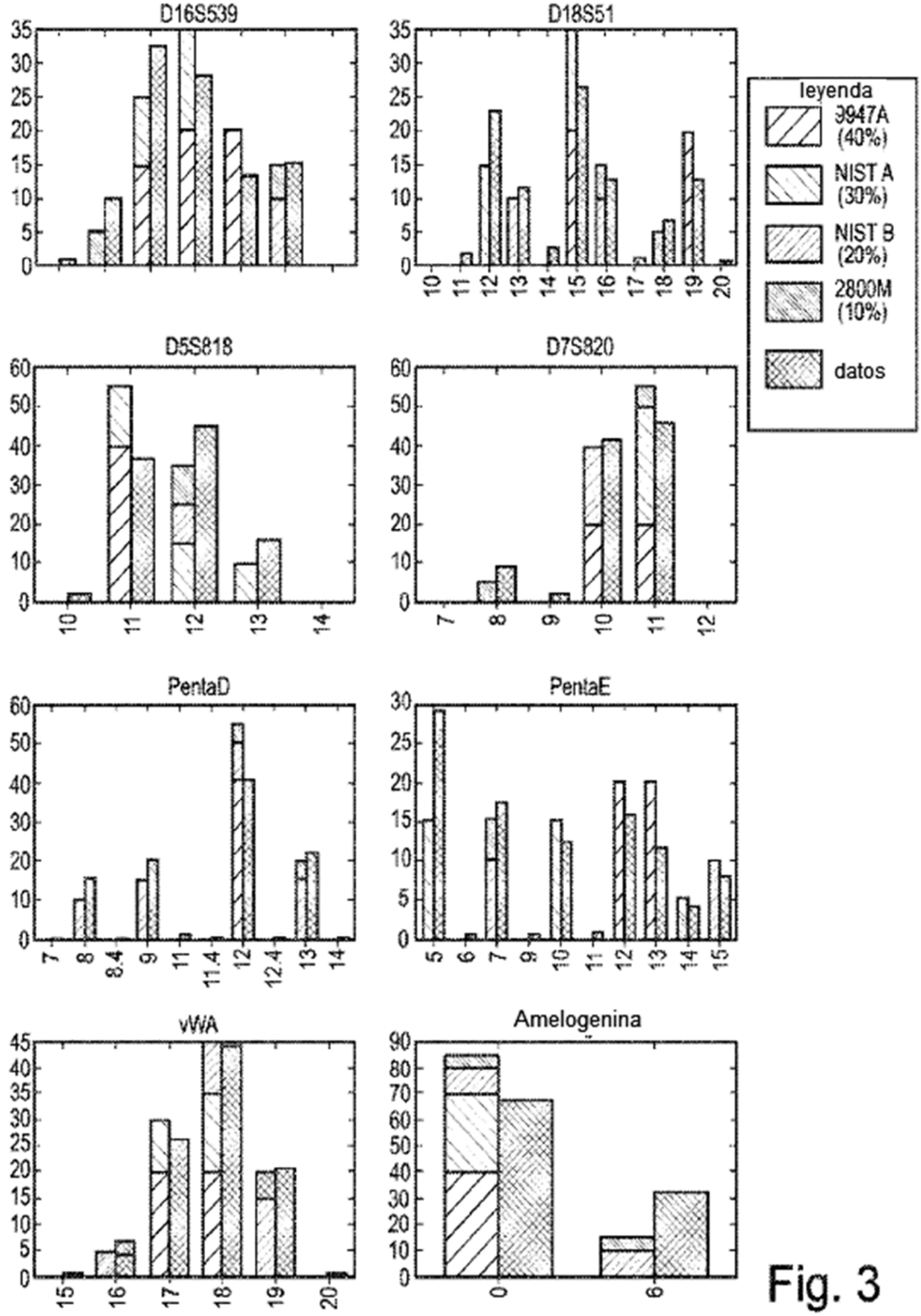

D16S539
D18S51
D5S818
D7S820
PentaD
PentaE
vWA
Amelogenina
leyenda
3947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
datos

Fig. 3

Muestra ADN	CSF1PO	D3S1358	D7S820	D16S539	D18S51	FGA	PentaE	TH01	vWA	D5S818	D8S1179	D13S317	D21S11	PentaD	TPOX
9947A	10, 12	14, 15	10, 11	11, 12	15, 19	23, 24	12, 13	8, 9.3	19, 20	11	13, 13	11	30	12	8
2800M	12	17, 18	8, 11	9, 13	16, 18	20, 23	7, 14	6, 9.3	16, 19	12	14, 15	9, 11	29, 31.2	12, 13	11
NIST A	10	15, 16	11	10, 11	12, 15	21, 23	5, 10	8, 9.3	18, 19	11, 12	13, 14	8	28, 32.2	9, 13	8
NIST B	10, 11	15, 19	10	10, 13	13, 16	20, 23	7, 15	6, 9.3	17, 18	12, 13	10, 13	9, 12	32, 32.2	8, 12	8, 11
NIST C	10, 12	16, 18	10, 12	10	16, 19	24, 26	12, 13	6, 8	16, 18	10, 11	10, 17	11	29, 30	10, 11	11
	SNP descubierto en mitad de las repeticiones: 46% [TCTA] 13,56% [TCTA]1 [TCTG]1 [TCTA]11														

Fig. 4