(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 625 288**

(51) Int. CI.:

C12Q 1/68 (2006.01)
C12N 15/11 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **12.04.2012 PCT/US2012/033207**

(87) Fecha y número de publicación internacional: **18.10.2012 WO12142213**

(96) Fecha de presentación y número de la solicitud europea: **12.04.2012 E 12772013 (4)**

(97) Fecha y número de publicación de la concesión europea: **05.04.2017 EP 2697397**

(54) Título: **Sistema de secuenciación segura**

(30) Prioridad:

15.04.2011 US 201161476150 P
10.05.2011 US 201161484482 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
19.07.2017

(73) Titular/es:

THE JOHNS HOPKINS UNIVERSITY (100.0%)
100 N. Charles Street 5th Floor
Baltimore, MD 21201, US

(72) Inventor/es:

VOGELSTEIN, BERT;
KINZLER, KENNETH W.;
PAPADOPOULOS, NICKOLAS y
KINDE, ISAAC

(74) Agente/Representante:

DURÁN MOYA, Luis Alfonso

DESCRIPCIÓN

Sistema de secuenciación segura

SECTOR TÉCNICO DE LA PRESENTE INVENCIÓN

La presente invención se refiere al área de secuenciación de ácidos nucleicos. En particular, se refiere a etapas manipulativas y analíticas para analizar y verificar los productos de acontecimientos de frecuencia baja.

ANTECEDENTES DE LA PRESENTE INVENCIÓN

Las mutaciones genéticas son la base de muchos aspectos de la vida y la muerte, a lo largo de la evolución y la enfermedad, respectivamente. En consecuencia, su medición es crucial para varios campos de investigación. El análisis clásico de la fluctuación de Luria y Delbrück es un ejemplo prototípico de las ideas sobre los procesos biológicos que se pueden obtener simplemente contando el número de mutaciones en experimentos controlados cuidadosamente (1). El recuento de las mutaciones *de novo* en los seres humanos, no presentes en sus padres, han dado lugar, de forma similar, a nuevos conocimientos sobre la velocidad a la que nuestra especie puede evolucionar (2, 3). Del mismo modo, el recuento de los cambios genéticos o epigenéticos en los tumores puede dar información sobre cuestiones fundamentales en la biología del cáncer (4). Las mutaciones están en el centro de los problemas actuales en el manejo de pacientes con enfermedades víricas, tales como el SIDA y la hepatitis, en virtud de la resistencia a los fármacos que pueden causar (5, 6). La detección de tales mutaciones, particularmente en una etapa previa a su predominio en la población, será, probablemente, esencial para optimizar la terapia. La detección del ADN del donante en la sangre de los pacientes con trasplante de órganos es un indicador importante del rechazo del injerto y la detección del ADN fetal en el plasma materno se puede utilizar para el diagnóstico prenatal de una manera no invasiva (7,8). En las enfermedades neoplásicas, que están todas impulsadas por mutaciones somáticas, las aplicaciones de la detección de mutantes raros son múltiples; pueden utilizarse para ayudar a identificar la enfermedad residual en los márgenes quirúrgicos o en los ganglios linfáticos, para seguir el curso de la terapia cuando se evalúan en el plasma y, tal vez, para identificar pacientes con enfermedad temprana curable con cirugía cuando se evalúa en heces, esputo, plasma y otros fluidos corporales (9-11).

Estos ejemplos destacan la importancia de identificar mutaciones raras tanto para la investigación básica como la clínica. En consecuencia, a lo largo de los años se han ideado maneras innovadoras de evaluarlos Los primeros procedimientos implicaron ensayos biológicos basados en prototrofia, resistencia a infecciones víricas o fármacos, o ensayos bioquímicos (1, 12-18). La clonación molecular y la secuenciación proporcionaron una nueva dimensión al sector, ya que permitieron identificar el tipo de mutación, más que su simple presencia (19-24). Algunos de los más potentes de estos nuevos procedimientos se basan en la PCR digital, en la que las moléculas individuales se evalúan una por una (25). La PCR digital es conceptualmente idéntica al análisis de clones individuales de bacterias, células o virus, pero se realiza completamente *in vitro* con reactivos inanimados definidos. Se han descrito varias implementaciones de la PCR digital, incluyendo el análisis de moléculas dispuestas en placas de múltiples pocillos, en polonias, en dispositivos microfluídicos y en emulsiones de agua en aceite (25-30). En cada una de estas tecnologías, los moldes mutantes se identifican mediante su unión a oligonucleótidos específicos para la base potencialmente mutante.

La secuenciación masiva paralela representa una forma particularmente potente de PCR digital en la que cientos de millones de moldes de moléculas pueden analizarse una por una. Tiene la ventaja sobre los procedimientos convencionales de PCR digital de que se pueden consultar muchas bases secuencial y fácilmente de una manera automatizada. Sin embargo, la secuenciación masivamente paralela generalmente no puede utilizarse para detectar variantes raras debido a la alta tasa de error asociada con el proceso de secuenciación. Por ejemplo, con los instrumentos de secuenciación Illumina de utilización habitual, esta tasa de error varía del ~1% (31, 32) al ~0,05% (33, 34), dependiendo de factores, tales como la longitud de lectura (35), la utilización de algoritmos de base mejorada (36-38) y el tipo de variantes detectadas (39). Algunos de estos errores se deben, probablemente, a las mutaciones introducidas durante la preparación del molde, durante las etapas de preamplificación requeridas para la preparación de la biblioteca y durante la posterior amplificación en fase sólida sobre el propio instrumento. Otros errores se deben a la mala incorporación de la base durante la secuenciación y los errores de identificación de bases. Los avances en la identificación de bases pueden aumentar la confianza (por ejemplo, (36-39)), pero los errores basados en instrumentos todavía son limitantes, particularmente en muestras clínicas en las que la prevalencia de la mutación puede ser del 0,01% o menos (11). En el trabajo que se describe a continuación, los presentes inventores muestran cómo se pueden preparar los moldes y los datos de secuenciación obtenidos de ellos pueden interpretarse con mayor fiabilidad, de modo que se puedan identificar mutaciones relativamente raras con instrumentos disponibles comercialmente.

Existe una necesidad continua en la técnica de mejorar la sensibilidad y exactitud de las determinaciones de secuencias para fines de investigación, clínicos, forenses y genealógicos.

CARACTERÍSTICAS DE LA PRESENTE INVENCIÓN

De acuerdo con un aspecto de la presente invención, un procedimiento analiza secuencias de ácido nucleico. Una secuencia de ácido nucleico de identificación única (UID) se une a un primer extremo de cada uno de una pluralidad de fragmentos de ácido nucleico de analito para formar fragmentos de ácido nucleico de analito identificados de forma única. La secuencia de nucleótidos de un fragmento de ácido nucleico de analito identificado de forma única se determina de forma redundante, en la que determinadas secuencias de nucleótidos que comparten una UID forman una familia de miembros. Se identifica una secuencia de nucleótidos que representa con precisión un fragmento de ácido nucleico de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y la secuencia se encuentra, como mínimo, en dos familias.

De acuerdo con otro aspecto de la presente invención, un procedimiento analiza secuencias de ácido nucleico. Una secuencia identificadora única (UID) está unida a un primer extremo de cada uno de una pluralidad de fragmentos de ADN de analito utilizando, como mínimo, dos ciclos de amplificación con el primer y segundo cebadores para formar fragmentos de ADN de analito identificados de forma única. Las UID están en exceso de los fragmentos de ADN de analito durante la amplificación. Los primeros cebadores comprenden un primer segmento complementario a un amplicón deseado; un segundo segmento que contiene la UID; y un tercer segmento que contiene un sitio de cebado universal para su posterior amplificación. Los segundos cebadores comprenden un sitio de cebado universal para la posterior amplificación. Cada ciclo de amplificación une un sitio de cebado universal a una hebra. Los fragmentos de ADN de analito identificados de forma única se amplifican para formar una familia de fragmentos de ADN de analito identificados de forma única a partir de cada fragmento de ADN de analito identificado de forma única. Se determinan secuencias de nucleótidos de una pluralidad de miembros de la familia.

Otro aspecto de la presente invención es un procedimiento para analizar ADN utilizando secuencias identificadoras únicas (UID) endógenas. Se obtiene ADN de analito fragmentado que comprende fragmentos de 30 a 2.000 bases, ambas incluidas. Cada extremo de un fragmento forma una UID endógena para el fragmento. Los oligonucleótidos adaptadores se unen a los extremos de los fragmentos para formar fragmentos adaptados. Los fragmentos que representan uno o más genes seleccionados se enriquecen opcionalmente mediante la captura de un subconjunto de los fragmentos utilizando oligonucleótidos de captura complementarios a genes seleccionados en el ADN de analito o amplificando fragmentos complementarios a genes seleccionados. Los fragmentos adaptados se amplifican utilizando cebadores complementarios a los oligonucleótidos adaptadores para formar familias de fragmentos adaptados. Se determina la secuencia de nucleótidos de una pluralidad de miembros de una familia. Se comparan secuencias de nucleótidos de la pluralidad de miembros de la familia. Se identifica una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 1% de los miembros de la familia contiene la secuencia.

Todavía otro aspecto de la presente invención es una composición que comprende la población de pares de cebadores, en la que cada par comprende un primer cebador y un segundo cebador para amplificar e identificar un gen o parte de un gen. El primer cebador comprende una primera parte de 10-100 nucleótidos complementarios al gen o parte del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un tercer cebador. El segundo cebador comprende una primera parte de 10-100 nucleótidos complementarios al gen o parte del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un cuarto cebador. Interpuesta entre la primera parte y la segunda parte del segundo cebador hay una tercera parte que consiste en de 2 a 4.000 nucleótidos que forman un identificador único (UID). Los identificadores únicos en la población tienen, como mínimo, 4 secuencias diferentes. El primer y segundo cebadores son complementarios a hebras opuestas al gen o parte de gen. Un kit puede comprender la población de cebadores y el tercer y cuarto cebadores complementarios a las segundas partes de cada uno del primer y segundo cebadores.

Estas y otras realizaciones que serán evidentes para los expertos en la materia tras la lectura de la memoria descriptiva dan a conocer la técnica herramientas y procedimientos para determinar de manera sensible y precisa las características o secuencias de los ácidos nucleicos.

DESCRIPCIÓN BREVE DE LOS DIBUJOS

Figura 1. **Elementos esenciales de Safe–SeqS.** En la primera etapa, a cada fragmento que se va a analizar se le asigna una secuencia identificadora única (UID) (barras rayadas o punteadas de metal). En el segundo paso, los fragmentos marcados de forma única se amplifican, produciendo familias de UID, cada miembro de las cuales tiene la misma UID. Un supermutante se define como una familia de UID en la que ≥ 95% de los miembros de la familia tiene la misma mutación.

Figura 2. **Safe–SeqS con UID endógenos más captura.** Las secuencias de los extremos de cada fragmento producido mediante cizallamiento aleatorio (barras sombreadas variables) sirven como identificadores únicos (UID). Estos fragmentos se ligan a adaptadores (barras sombreadas con cuadrados y rayas cruzadas) para que puedan amplificarse posteriormente mediante PCR. Se produce un fragmento identificable de manera única a partir de cada hebra del molde bicatenario; solo se muestra una hebra. Los fragmentos de interés se capturan en una fase sólida que contiene oligonucleótidos complementarios a las secuencias de interés. Después de la amplificación mediante PCR para producir familias de UID con cebadores que contienen secuencias “de injerto” en 5' (barras sombreadas

con rayas inclinadas y puntos separados), se lleva a cabo la secuenciación y los supermutantes se definen, tal como en la figura 1.

Figura 3. **Safe–SeqS con UID exógenas.** El ADN (cortado o sin cortar) se amplifica con un conjunto de cebadores específicos del gen. Uno de los cebadores tiene una secuencia de ADN aleatoria (*por ejemplo,* un conjunto de 14 N) que forma el identificador único (UID, barras variables), situado en 5' de la secuencia específica de su gen, y ambos tienen secuencias que permiten la amplificación universal en la siguiente etapa (barras con cuadrados y rayas cruzadas). Dos ciclos de asignación de UID producen dos fragmentos, cada uno con una UID diferente, de cada molécula molde bicatenaria, tal como se muestra. La PCR posterior con cebadores universales, que también contienen secuencias de "injerto" (barras sombreadas con rayas inclinadas y puntos separados), produce familias de UID que están secuenciadas directamente. Los supermutantes se definen como en la leyenda de la figura 1.

Figuras 4A-4B. **Sustituciones de una sola base identificadas mediante análisis convencional y Safe–SeqS.** Se utilizó la estrategia de UID exógena representada en la figura 3 para producir fragmentos de PCR del gen *CTNNBI* de tres individuos normales no relacionados. Cada posición representa una de las 87 posibles sustituciones de una sola base (analizadas 3 posibles sustituciones/base x 29 bases). Estos fragmentos se secuenciaron en un instrumento Illumina GA IIx y se analizaron de la manera convencional (figura 4A) o con Safe–SeqS (figura 4B). Los resultados de Safe–SeqS se muestran en la misma escala que el análisis convencional para la comparación directa; el recuadro insertado es una vista ampliada. Obsérvese que la mayoría de las variantes identificadas mediante el análisis convencional es probable que representen errores de secuenciación, como lo indica su alta frecuencia relativa a Safe–SeqS y su coherencia entre las muestras no relacionadas.

Figura 5. **Safe–SeqS con UDI endógenas más PCR inversa.** La secuencia de los extremos de cada fragmento producido mediante cizallamiento aleatorio sirve como identificadores únicos (UID; barras sombreadas variables). Estos fragmentos se ligan a los adaptadores (barras con cuadrados y rayas cruzadas) como en una preparación de biblioteca estándar de Illumina. Se produce un fragmento marcado de forma única a partir de cada hebra del molde bicatenario; solo se muestra una hebra. Después de la circularización con una ligasa, se realiza PCR inversa con cebadores específicos de genes que también contienen secuencias de "injerto" en 5' (barras sombreadas con rayas inclinadas y puntos separados). Esta PCR produce familias de UID que están secuenciadas directamente. Los supermutantes se definen como en la figura 1.

Figura 6A-6B. **Posición de las sustituciones de una sola base frente a la frecuencia de los errores en los oligonucleótidos sintetizados con fosforoamiditas y Phusion.** Una parte representativa del mismo fragmento de ADN de 31 bases sintetizado con fosforoamiditas (figura 6A) o la polimerasa Phusion (figura 6B) se analizó mediante Safe–SeqS. Se representan las medias y las desviaciones estándar para siete experimentos independientes de cada tipo. Se observó un promedio de 1.721 ± 383 y 196 ± 143 de supermutantes de SBS identificados en los fragmentos generados por Phusion y sintetizados con fosforoamidita, respectivamente. El eje y indica la fracción de los errores totales en la posición indicada. Obsérvese que los errores en el fragmento de ADN sintetizado con fosforoamidita fueron consistentes entre las siete repeticiones, como cabría esperar si los errores se introdujeran sistemáticamente durante la propia síntesis. Por el contrario, los errores en los fragmentos generados por Phusion parecían ser heterogéneos entre las muestras, como cabría esperar de un proceso estocástico (Luria y Delbruck, Genetics 28: 491–511, 1.943).

Figura 7. Distribución del miembro de la familia de UID. La estrategia de UID exógena representada en la figura 3 se utilizó para producir fragmentos de PCR a partir de una región de *CTNNBI* de tres individuos normales no relacionados (tabla 2B); se muestra un ejemplo representativo de las familias de UID con ≤ 300 miembros (el 99% del total de familias de UID) generadas a partir de un individuo. El eje y indica el número de familias de UID diferentes que contenían el número de miembros de la familia mostrados en el eje x.

DESCRIPCIÓN DETALLADA DE LA PRESENTE INVENCIÓN

Los presentes inventores han desarrollado un enfoque, denominado "Safe-SeqS" (de *Safe–Sequencing System*, sistema de secuenciación segura). En una realización, implica dos etapas básicas (figura 1). La primera es la asignación de un identificador único (UID) a cada molécula de molde de ácido nucleico que se va a analizar. La segunda es la amplificación de cada molde marcado de forma única, de manera que se generan muchas moléculas hija con la secuencia idéntica (definida como una familia de UID). Si una mutación existiera previamente en la molécula molde utilizada para la amplificación, esa mutación debería estar presente en una determinada proporción, o incluso en todas, de las moléculas hija que contienen dicho UID (salvo cualquier posterior replicación o errores de secuenciación). Una familia de UID en la cual cada miembro de la familia (o una cierta proporción predeterminada) tiene una mutación idéntica se llama un "supermutante". Las mutaciones que no se producen en los moldes originales, tales como las que se producen durante las etapas de amplificación o mediante errores en la identificación de bases, no deberían dar lugar a supermutantes, es decir, no estarán presentes en la frecuencia predeterminada en una familia de UID. En otras realizaciones, la amplificación no es necesaria.

El enfoque se puede utilizar para cualquier propósito en el que se requiera un nivel muy alto de precisión y sensibilidad a partir de los datos de secuencia. Tal como se muestra a continuación, el enfoque puede utilizarse para evaluar la fidelidad de una polimerasa, la precisión de la síntesis de ácidos nucleicos sintetizados *in vitro* y la

prevalencia de mutaciones en ácidos nucleicos nucleares o mitocondriales de células normales. El enfoque puede utilizarse para detectar y/o cuantificar mosaicismo y mutaciones somáticas.

Pueden obtenerse fragmentos de ácidos nucleicos utilizando una técnica de formación de fragmentos aleatoria tal como cizallamiento mecánico, sonicación o sometiendo los ácidos nucleicos a otras tensiones físicas o químicas. Los fragmentos pueden no ser estrictamente aleatorios, ya que algunos sitios pueden ser más susceptibles a las tensiones que otros. También se pueden utilizar endonucleasas que se fragmentan aleatoriamente o específicamente para generar fragmentos. El tamaño de los fragmentos puede variar, pero de forma deseable estará en intervalos entre 30 y 5.000 pares de bases, entre 100 y 2.000, entre 150 y 1.000, o dentro de intervalos con diferentes combinaciones de estos criterios de valoración. Los ácidos nucleicos pueden ser, por ejemplo, ARN o ADN. También se pueden utilizar formas modificadas de ARN o ADN.

La unión de un UID exógeno a un fragmento de ácidos nucleicos de analito puede realizarse por cualquier medio conocido en la técnica, incluyendo enzimático, químico o biológico. Un medio utiliza una reacción en cadena de la polimerasa. Otro medio utiliza una enzima ligasa. La enzima puede ser, por ejemplo, de mamífero o bacteriana. Los extremos de los fragmentos pueden repararse antes de la unión utilizando otras enzimas, tales como el fragmento de Klenow de la ADN polimerasa de T4. Otras enzimas que pueden utilizarse para la unión son otras enzimas polimerasas. Se puede añadir un UID a uno o ambos extremos de los fragmentos. Un UID puede estar contenido dentro de una molécula de ácido nucleico que contiene otras regiones para otra funcionalidad deseada. Por ejemplo, se puede añadir un sitio de cebado universal para permitir una amplificación posterior. Otro sitio adicional puede ser una región de complementariedad con una región o gen particular en los ácidos nucleicos de analito. Un UID puede tener una longitud de 2 a 4.000, de 100 a 1.000, de 4 a 400, bases, por ejemplo.

Los UID pueden hacerse utilizando adición aleatoria de nucleótidos para formar una secuencia corta que se utilizará como identificador. En cada posición de adición, se puede utilizar una selección de uno de los cuatro desoxirribonucleótidos. Como alternativa, se puede utilizar una selección de uno de tres, dos o un desoxirribonucleótido. Por tanto, el UID puede ser completamente aleatorio, algo aleatorio o no aleatorio en ciertas posiciones. Otra forma de hacer UID utiliza nucleótidos predeterminados montados en un chip. En esta manera, la complejidad se alcanza de una manera planificada. Puede ser ventajoso unir un UID a cada extremo de un fragmento, aumentando la complejidad de la población de UID en fragmentos.

Un ciclo de la reacción en cadena de la polimerasa para añadir UID exógeno se refiere a la desnaturalización térmica de una molécula bicatenaria, a la hibridación de un primer cebador a una hebra simple resultante, a la extensión del cebador para formar una segunda hebra nueva hibridada a la hebra única original. Un segundo ciclo se refiere a la desnaturalización de la segunda hebra nueva de la hebra única original, a la hibridación de un segundo cebador a la nueva segunda hebra nueva y a la extensión del segundo cebador para formar una tercera hebra nueva, hibridada a la nueva segunda hebra. Pueden requerirse múltiples ciclos para aumentar la eficiencia, por ejemplo, cuando el analito está diluido o hay inhibidores presentes.

En el caso de los UID endógenos, se pueden añadir adaptadores a los extremos de fragmentos mediante ligación. La complejidad de los fragmentos de analito puede disminuirse mediante una etapa de captura, ya sea en una fase sólida o en una etapa líquida. Típicamente, la etapa de captura utilizará hibridación con sondas que representan un gen o conjunto de genes de interés. En caso de una fase sólida, los fragmentos que no son de unión se separan de los fragmentos de unión. Las fases sólidas adecuadas conocidas en la materia incluyen filtros, membranas, perlas, columnas, etc. En caso de una fase líquida, se puede añadir un reactivo de captura que se une a las sondas, por ejemplo a través de una interacción de tipo biotina-avidina. Después de la captura, los fragmentos deseados se pueden eluir para su posterior procesamiento. El orden de la adición de adaptadores y la captura no es crítico. Otro medio para reducir la complejidad de los fragmentos de analito implica la amplificación de uno o más genes o regiones específicos. Una forma de lograr esto es utilizar la PCR inversa. Pueden utilizarse cebadores que son específicos de los genes, enriqueciendo al mismo tiempo la formación de bibliotecas. Opcionalmente, los cebadores específicos de gen pueden contener secuencias de injerto para su posterior unión a una plataforma de secuenciación masivamente paralela.

Debido a que los UID endógenos proporcionan un número limitado de posibilidades únicas, dependiendo del tamaño del fragmento y la longitud de la lectura de la secuenciación, pueden utilizarse combinaciones de UID tanto endógenos como exógenos. La introducción de secuencias adicionales al amplificar aumentaría los UID disponibles y, de este modo, aumentaría la sensibilidad. Por ejemplo, antes de la amplificación, el molde puede dividirse en 96 pocillos y se podrían utilizar 96 cebadores diferentes durante la amplificación. Esto aumentaría con eficacia los UID disponibles 96 veces, porque se podrían distinguir hasta 96 moldes con el mismo UID endógeno. Esta técnica también puede utilizarse con UID exógenos, de manera que los cebadores de cada pocillo añaden una secuencia única, bien específica, a los productos de amplificación. Esto puede mejorar la especificidad de la detección de moldes raros.

La amplificación de fragmentos que contienen un UID se puede realizar de acuerdo con técnicas conocidas para generar familias de fragmentos. Se puede utilizar la reacción en cadena de la polimerasa. También pueden utilizarse otros procedimientos de amplificación, tal como como sea conveniente. Se puede utilizar la PCR inversa, al igual

que la amplificación por círculo rodante. La amplificación de fragmentos típicamente se realiza utilizando cebadores que son complementarios a los sitios de cebado que están unidos a los fragmentos al mismo tiempo que los UID. Los sitios de cebado son distales a los UID, de modo que la amplificación incluye los UID. La amplificación forma una familia de fragmentos, de modo que cada miembro de la familia comparte el mismo UID. Debido a que la diversidad de UID es, en gran medida, superior a la diversidad de los fragmentos, cada familia debe derivar de un único fragmento de molécula en el analito. Los cebadores utilizados para la amplificación pueden modificarse químicamente para hacerlos más resistentes a las exonucleasas. Una de tales modificaciones es la utilización de enlaces fosforotioato entre uno o más nucleótidos en 3'. Otro utiliza boranofosfatos.

Los miembros de la familia se secuencian y comparan para identificar cualquier divergencia dentro de una familia. La secuenciación se realiza, preferentemente, en una plataforma de secuenciación masivamente paralela, muchas de las cuales están disponibles comercialmente. Si la plataforma de secuenciación requiere una secuencia para "injerto", es decir, la unión al dispositivo de secuenciación, dicha secuencia se puede añadir durante la adición de UID o adaptadores o por separado. Una secuencia de injerto puede ser parte de un cebador de UID, un cebador universal, un cebador específico para una diana génica, los cebadores de amplificación utilizados para formar una familia o separados. La secuenciación redundante se refiere a la secuenciación de una pluralidad de miembros de una única familia.

Se puede establecer un umbral para identificar una mutación en un analito. Si la "mutación" aparece en todos los miembros de una familia, deriva de analito. Si aparece en menos de todos los miembros, puede haberse introducido durante el análisis. Los umbrales para identificar una mutación se pueden establecer, por ejemplo, al 50%, 60%, 70%, 80%, 90%, 95%, 97%, 98% o 100%. Los umbrales se establecerán sobre la base del número de miembros de una familia que están secuenciados y el propósito y la situación en particular.

Las poblaciones de pares de cebadores se utilizan para unir UID exógenos. El primer cebador comprende una primera parte de 10-100 nucleótidos complementarios al gen o parte del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un tercer cebador. El segundo cebador comprende una primera parte de 10-100 nucleótidos complementarios al gen o parte del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un cuarto cebador. Interpuesta entre la primera parte y la segunda parte del segundo cebador hay una tercera parte que consiste en de 2 a 4.000 nucleótidos que forman un identificador único (UID). Los identificadores únicos en la población tienen, como mínimo, 4, como mínimo 16, como mínimo 64, como mínimo 256, como mínimo 1.024, como mínimo 4.096, como mínimo 16.384, como mínimo 65.536, como mínimo 262.144, como mínimo 1.048.576, como mínimo 4.194.304, como mínimo 16.777.216 o, como mínimo, 67.108.864 secuencias diferentes. El primer y segundo cebadores son complementarios a hebras opuestas del gen o parte de gen. Se puede hacer un kit que contenga ambos cebadores para unir los UID exógenos, así como cebadores de amplificación, es decir, el tercer y cuarto cebadores complementarios a las segundas partes de cada uno de los cebadores primero y segundo. El tercer y cuarto cebadores pueden contener, opcionalmente, secuencias de injerto o indexación adicionales. El UID puede comprender secuencias seleccionadas aleatoriamente, secuencias de nucleótidos predefinidas, o ambas secuencias seleccionadas aleatoriamente y nucleótidos predefinidos. En los dos casos, estas pueden estar unidas en bloques o entremezcladas.

Los procedimientos de análisis se pueden utilizar para cuantificar, así como para determinar una secuencia. Por ejemplo, se puede comparar la abundancia relativa de dos fragmentos de ADN de analito.

Los resultados descritos a continuación en los ejemplos demuestran que el enfoque Safe-SeqS puede mejorar sustancialmente la precisión de la secuenciación masivamente paralela (tablas 1 y 2). Puede implementarse a través de UID endógenos o introducidos exógenamente y puede aplicarse a prácticamente cualquier flujo de trabajo de preparación de muestras o plataforma de secuenciación. Tal como se demuestra en el presente documento, el enfoque puede utilizarse fácilmente para identificar mutantes raros en una población de moldes de ADN, medir las tasas de error de la polimerasa y juzgar la fiabilidad de las síntesis de oligonucleótidos. Una de las ventajas de la estrategia es que produce el número de moldes analizados, así como la fracción de moldes que contienen bases variantes. Los procedimientos descritos previamente *in vitro* para la detección de un pequeño número de moléculas molde (por ejemplo, (29, 50)) permiten determinar la fracción de moldes mutantes pero no pueden determinar el número de moldes mutantes y normales en la muestra original.

Es de interés comparar la Safe-SeqS con otros enfoques para reducir errores en la secuenciación de próxima generación. Tal como se ha mencionado anteriormente, en los antecedentes de la presente invención, se han desarrollado algoritmos sofisticados para aumentar la precisión de la identificación de bases (por ejemplo, (36-39)). Estos pueden reducir, ciertamente, las identificaciones de falsos positivos, pero su sensibilidad aún está limitada por mutaciones artifactuales que se producen durante las etapas PCR requeridas para la preparación de la biblioteca así como por (un número reducido de) errores de identificación de bases. Por ejemplo, el algoritmo utilizado en el estudio actual utilizó criterios muy estrictos para la identificación de bases y se aplicó a longitudes de lectura corta, pero todavía no pudo reducir la tasa de error a menos de un promedio de $2,0 \times 10^{-4}$ errores/pb. Esta frecuencia de error es, como mínimo, tan baja como la indicada con otros algoritmos. Para mejorar aún más la sensibilidad, estas mejoras de la identificación de bases se pueden utilizar junto con Safe–SeqS. Travers y otros han descrito otra potente estrategia para reducir los errores (51). Con esta tecnología, ambas cadenas de cada molécula molde se

secuencian de forma redundante después de una serie de etapas enzimáticas preparativas. Sin embargo, este enfoque solo puede realizarse en un instrumento específico. Además, para muchas aplicaciones clínicas, hay relativamente pocas moléculas molde en la muestra inicial y se requiere una evaluación de casi todas ellas para obtener la sensibilidad requerida. El enfoque descrito en el presente documento con UID introducidos de forma exógena (figura 3) cumple este requisito acoplando la etapa de asignación de UID con una amplificación posterior en la que se pierden pocas moléculas. Nuestros enfoques de UID endógenos (figura 2 y figura 5) y el descrito por Travers y otros no son ideales para este propósito debido a las inevitables pérdidas de moléculas molde durante la ligación y otras etapas preparativas.

¿Cómo se sabe que las mutaciones identificadas mediante los análisis convencionales en el estudio actual representan artefactos en lugar de mutaciones verdaderas en los moldes originales? Muchas pruebas avalan la observación de que la prevalencia de la mutación en todos los experimentos excepto en uno era similar, de 2,0 x 10^{-4} a 2,4 x 10^{-4} mutaciones/pb (tablas 1 y 2). La excepción fue el experimento con oligonucleótidos sintetizados a partir de fosforoamiditas, en los que el error del proceso de síntesis era aparentemente más alto que la tasa de error del análisis de Illumina convencional cuando se usó con criterios de identificación de bases rigurosos. En contraste, la prevalencia de la mutación de Safe-SeqS varió mucho más, de 0,0 a 1,4 x 10^{-5} mutaciones/pb, dependiendo del molde y el experimento. Además, la prevalencia de la mutación medida mediante Safe-SeqS en el experimento más controlado, en el que se midió la fidelidad de la polimerasa (tabla 2A), fue casi idéntica a la predicha a partir de experimentos previos en los que se midió la fidelidad de la polimerasa mediante ensayos biológicos. Las mediciones de los presentes inventores de la prevalencia de la mutación en el ADN de las células normales son coherentes con algunos datos experimentales anteriores. Sin embargo, las estimaciones de estas prevalencias varían ampliamente y pueden depender del tipo de célula y la secuencia analizada (véase el texto SI). Por lo tanto, los presentes inventores no pueden estar seguros de que las pocas mutaciones reveladas por Safe-SeqS representaban errores que se producen durante el proceso de secuenciación en lugar de mutaciones verdaderas presentes en los moldes de ADN original. Las posibles fuentes de error en el proceso Safe-SeqS se describen en el texto SI.

Otra aplicación potencial de Safe-SeqS es la minimización de la contaminación de la PCR, un serio problema para los laboratorios clínicos. Con asignación de UID endógeno o exógeno, los UID de moldes mutantes pueden compararse simplemente con los identificados en experimentos previos; la probabilidad de que la misma mutación de dos muestras independientes tuviera el mismo UID en diferentes experimentos es insignificante cuando las mutaciones son infrecuentes. Además, con UID exógenos, un experimento de control con el mismo molde pero sin el UID que asigna ciclos de PCR (figura 3) puede asegurar que no hay contaminación de ADN presente en dicha preparación de molde; ningún molde debe amplificarse en ausencia de ciclos de asignación de UID y, por lo tanto, no debe observarse ningún producto de PCR del tamaño adecuado.

Como todas las técnicas, Safe-SeqS tiene limitaciones. Por ejemplo, los presentes inventores han demostrado que puede utilizarse la estrategia de UID exógeno para analizar un solo amplicón en profundidad. Esta tecnología puede no ser aplicable a situaciones en las que se deben analizar múltiples amplicones a partir de una muestra que contiene un número limitado de moldes. La multiplexación en los ciclos de asignación de UID (figura 3) puede proporcionar una solución a este reto. Una segunda limitación es que la eficiencia de la amplificación en los ciclos de asignación de UID es crítica para el éxito del procedimiento. Las muestras clínicas pueden contener inhibidores que reducen la eficacia de esta etapa. Este problema se puede superar, presumiblemente, realizando más de dos ciclos en la etapa de PCR de asignación de UID (figura 3), aunque esto complicaría la determinación del número de moldes analizados. La especificidad de Safe-SeqS está actualmente limitada por la fidelidad de la polimerasa utilizada en la etapa de PCR de asignación de UID, es decir, 8,8 x 10^{-7} mutaciones/pb en su implementación actual con dos ciclos. Aumentar el número de ciclos en la etapa de PCR de asignación de UID a cinco disminuiría la especificidad general a ~2 x 10^{-6} mutaciones/pb. Sin embargo, esta especificidad puede incrementarse al requerir más de un supermutante para la identificación de la mutación, la probabilidad de introducir la misma mutación artifactual dos o tres veces sería excesivamente baja ($[2 \times 10^{-6}]^2$ o $[2 \times 10^{-6}]^3$, respectivamente). En resumen, existen varias maneras sencillas de realizar variaciones de Safe-Seqs y de análisis para realizar las necesidades de experimentos específicos.

Luria y Delbruck, en su artículo clásico de 1.943, escribieron que su "predicción no puede verificarse directamente, porque lo que observamos, cuando contamos el número de bacterias resistentes en un cultivo, no es el número de mutaciones que se han producido, sino el número de bacterias resistentes que han surgido por la multiplicación de las que han mutado, dependiendo la cantidad de la multiplicación de cuando se produjo la mutación." El procedimiento Safe-SeqS descrito en el presente documento puede verificar tales predicciones porque el número, así como el tiempo de ocurrencia de cada mutación, se pueden estimar a partir de los datos, tal como se observa en los experimentos sobre la fidelidad de la polimerasa. Además de los moldes generados por las polimerasas in vitro, se puede aplicar el mismo enfoque al ADN de bacterias, virus y células de mamíferos. Por lo tanto, cabe esperar que esta estrategia proporcione respuestas definitivas a diversas cuestiones biomédicas importantes.

La divulgación anterior describe, en general, la presente invención. Se puede obtener un entendimiento más completo haciendo referencia a los siguientes ejemplos específicos que se proporcionan en el presente documento solo con fines ilustrativos y no pretenden limitar el alcance de la presente invención definida en las reivindicaciones de la presente invención.

EJEMPLO 1 – UID endógenos

Los UID, a veces llamados códigos de barras o índices, pueden asignarse a fragmentos de ácido nucleico de muchas maneras. Estos incluyen la introducción de secuencias exógenas mediante PCR (40, 41) o ligadura (42, 43). Todavía de un modo más simple, el ADN genómico cortado al azar contiene inherentemente UID que consisten en las secuencias de los dos extremos de cada fragmento cortado (figura 2 y figura 5). La secuenciación de extremos pareados de estos fragmentos produce familias de UID que se pueden analizar, tal como se ha descrito anteriormente. Para utilizar estos UID endógenos en Safe-SeqS, se utilizaron dos enfoques separados: uno diseñado para evaluar muchos genes de forma simultánea y el otro diseñado para evaluar un fragmento de un solo gen en profundidad (figura 2 y figura 5, respectivamente).

Para la evaluación de múltiples genes, los presentes inventores ligaron adaptadores de secuenciación Illumina estándar a los extremos de los fragmentos de ADN cortados para producir una biblioteca de secuenciación estándar, a continuación, capturaron los genes de interés en una fase sólida (44). En este experimento, se utilizó una biblioteca hecha a partir del ADN de ~15.000 células normales y se seleccionaron 2.594 pb de seis genes para la captura. Tras excluir los polimorfismos de nucleótido único conocidos, también se identificaron 25.563 mutaciones aparentes, correspondientes a $2{,}4 \times 10^{-4}$ ± mutaciones/pb (tabla 1). Basándose en análisis previos de las tasas de mutación en células humanas, era probable que, como mínimo, el 90% de estas mutaciones aparentes representaran mutaciones introducidas durante la preparación del molde y de la biblioteca o errores de identificación de bases. Obsérvese que la tasa de error determinada en el presente documento ($2{,}4 \times 10^{-4}$ mutaciones/pb) es considerablemente más baja de lo que normalmente se notifica en experimentos que utilizan el instrumento Illumina porque los presentes inventores utilizaron criterios muy estrictos para la identificación de bases.

Tabla 1. Safe–SeqS con UID endógenos

Análisis convencional	**Captura**	**PCR inversa**
Pb de alta calidad	106.958.863	1.041.346.645
Profundidad media de lectura de pb de alta calidad	38.620×	2.085.600×
Mutaciones identificadas	25.563	234.352
Mutaciones/pb	**2,4E–04**	**2,3E–04**
Análisis Safe–SeqS		
Pb de alta calidad	106.958.863	1.041.346.645
Profundidad media de lectura de pb de alta calidad	38.620×	2.085.600×
Familias de UID	69.505	1.057
Promedio de miembros/familia de UID	40	21.688
Mediana de miembros/familia de UID	19	4
Supermutantes identificados	8	0
Supermutantes/pb	**3,5E–06**	**0,0**

Con el análisis Safe-SeqS de los mismos datos, los presentes inventores determinaron que en este experimento se evaluaron 69.505 moléculas molde originales (es decir, se identificaron 69.505 familias de UID, con un promedio de 40 miembros por familia, tabla 1). Todas las variantes polimórficas identificadas mediante análisis convencional también se identificaron mediante Safe-SeqS. Sin embargo, solo se observaron 8 supermutantes entre estas familias, correspondientes a $3{,}5 \times 10^{-6}$ mutaciones/pb. Por lo tanto, Safe-SeqS disminuyó los posibles errores de secuenciación, como mínimo, 70 veces.

El análisis Safe-SeqS también puede determinar qué hebra de un molde está mutada, por lo que un criterio adicional para identificar mutaciones podría requerir que la mutación apareciese en solo una o en ambas cadenas del molde originalmente de doble cadena. Los secuenciadores masivamente paralelos son capaces de obtener información de secuencia de ambos extremos de un molde en dos lecturas secuenciales. (Este tipo de experimento de secuenciación se denomina "extremo apareado" en la plataforma Illumina, pero se pueden realizar experimentos similares en otras plataformas de secuenciación en las que pueden llamarse con otro nombre.) Las dos hebras de un molde de doble cadena pueden diferenciarse por la orientación observada de las secuencias y el orden en el que aparecen cuando se obtiene la información de secuencia de ambos extremos. Por ejemplo, un par de hebras de UID podría consistir en los dos grupos de secuencias siguientes cuando cada extremo de un molde se secuencia en lecturas secuenciales: 1) una secuencia en la orientación sentido que comienza en la posición 100 del cromosoma 2 en la primera lectura seguida de una secuencia en la orientación antisentido que comienza en la posición 400 del cromosoma 2 en la segunda lectura; y 2) una secuencia en la orientación antisentido que comienza en la posición 400 del cromosoma 2 en la primera lectura seguida de una secuencia en la orientación sentido que comienza en la posición 100 del cromosoma 2 en la segunda lectura. En el experimento de captura descrito anteriormente, 42.222 de 69.505 UID (que representan 21.111 moléculas de doble cadena originales) en la región de interés representaron

pares de hebras de UID. Estos 42.222 UID abarcaban 1.417.838 bases en la región de interés. Cuando se permita que una mutación solo se produjese dentro de pares de hebras de UID (ya sea en una o en ambas cadenas), se observaron dos supermutantes, que daban una tasa de mutación de 1,4 x 10^{-6} supermutantes/pb. Cuando se requirió que una mutación se produjera solo en una hebra de un par de hebras de UID, solo se observó un supermutante, dando una tasa de mutación de 7,1 x 10^{-7} supermutantes/pb. Cuando se requirió que una mutación se produjera en ambas hebras un par de hebras de UID, solo se observó un supermutante, dando una tasa de mutación de 7,1 x 10^{-7} supermutantes/pb. Por lo tanto, la exigencia de que las mutaciones se produzcan solo en una o en ambas hebras de los moldes puede aumentar aún más la especificidad de Safe-SeqS.

También se utilizó una estrategia que utiliza UID endógeno para reducir las mutaciones falsos positivos sobre secuenciación profunda de una sola región de interés. En este caso, se usó una biblioteca preparada, tal como se ha descrito anteriormente a partir de ~1.750 células normales como molde para PCR inversa utilizando cebadores complementarios a un gen de interés, de manera que los productos de PCR pudieron utilizarse directamente para la secuenciación (figura 5). Con el análisis convencional, se observó un promedio de 2,3 x 10^{-4} mutaciones/pb, similares a las observadas en el experimento de captura (tabla 1). Dado que en este experimento solo se evaluaron 1.057 moléculas independientes de las células normales, tal como se determinó mediante el análisis Safe-SeqS, todas las mutaciones observadas con el análisis convencional probablemente representaran falsos positivos (tabla 1). Con el análisis Safe-SeqS de los mismos datos no se identificaron supermutantes en ninguna posición.

EJEMPLO 2 – UID exógenos

Aunque los resultados descritos anteriormente muestran que Safe-SeqS puede aumentar la fiabilidad de la secuenciación masivamente paralela, el número de moléculas diferentes que pueden analizarse utilizando UID endógenos es limitado. Para los fragmentos cortados a un tamaño promedio de 150 pb (intervalo 125-175), la secuenciación de extremos apareados de 36 bases puede evaluar un máximo de ~7.200 moléculas diferentes que contienen una mutación específica (2 lecturas x 2 orientaciones x 36 bases/lectura x variación de 50 bases en cualquiera de los extremos del fragmento). En la práctica, el número real de UID es menor porque el proceso de corte no es totalmente aleatorio.

Para hacer una utilización más eficiente de los moldes originales, los presentes inventores desarrollaron una estrategia Safe-SeqS que utilizó un número mínimo de etapas enzimáticas. Esta estrategia también permitió la utilización de ADN degradado o dañado, tal como se encuentra en especímenes clínicos o después del tratamiento con bisulfito para el examen de la metilación de citosina (45). Tal como se representa en la figura 3, esta estrategia utiliza dos conjuntos de cebadores de PCR. El primer conjunto se sintetiza con precursores de fosforoamidita estándar y contenía secuencias complementarias al gen de interés en el extremo 3' y diferentes colas en los extremos 5' de ambos cebadores, directos e inversos. Las diferentes colas permitieron la amplificación universal en la siguiente etapa. Finalmente, hubo un tramo de 12 a 14 nucleótidos aleatorios entre la cola y los nucleótidos específicos de la secuencia en el cebador directo (40). Los nucleótidos aleatorios forman los UID. Una forma equivalente de asignar UID a fragmentos, no utilizada en este estudio, utilizaría 10.000 cebadores directos y 10.000 cebadores inversos sintetizados en una micromatriz. Cada uno de estos 20.000 cebadores tendrían cebadores específicos de genes en sus extremos en 3' y una de 10.000 secuencias de UID específicas, predeterminadas, que no se solapan en sus extremos 5', permitiendo 10^8 (es decir, $[10^4]^2$) posibles combinaciones de UID. En cualquier caso, se realizan dos ciclos de PCR con los cebadores y una polimerasa de alta fidelidad, produciendo un fragmento de ADN de doble cadena marcado de manera única de cada una de las dos cadenas de cada molécula molde original (figura 3). Los cebadores de asignación de UID no utilizados residuales se eliminan mediante digestión con una exonucleasa específica de una sola hebra, sin purificación adicional, y se añaden dos nuevos cebadores. Como alternativa o, además, de tal digestión, se puede utilizar una columna de sílice que retenga selectivamente fragmentos de mayor tamaño o se pueden utilizar perlas de inmovilización reversible en fase sólida (SPRI) en condiciones que retengan selectivamente fragmentos mayores para eliminar artefactos de amplificación más pequeños, no específicos. Esta purificación puede ayudar potencialmente a reducir la acumulación de cebador-dímero en etapas posteriores. Los nuevos cebadores, complementarios a las colas introducidas en los ciclos de asignación de UID, contienen secuencias de injerto en sus extremos 5', lo que permite la amplificación en fase sólida en el instrumento Illumina y residuos de fosforotioato en sus extremos 3' para hacerlos resistentes a cualquier exonucleasa restante. Después de 25 ciclos adicionales de PCR, los productos se cargan en el instrumento Illumina. Tal como se muestra a continuación, esta estrategia permitió a los presentes inventores evaluar la mayoría de los fragmentos de entrada y se utilizó para varios experimentos ilustrativos.

EJEMPLO 3 – Análisis de la fidelidad de la ADN polimerasa

La medición de las tasas de error de las polimerasas de ADN es esencial para su caracterización y dicta las situaciones en las que estas enzimas pueden utilizarse. Los presentes inventores eligieron medir la tasa de error de la polimerasa Phusion, ya que esta polimerasa tiene una de las frecuencias de error más bajas notificadas de cualquier enzima disponible comercialmente y, por lo tanto, plantea un desafío particular para un enfoque basado *in vitro*. En primer lugar, los presentes inventores amplificaron una única molécula molde de ADN humano, que comprende un segmento de un gen humano elegido de forma arbitraria, a través de 19 rondas de PCR. Los productos de la PCR de estas amplificaciones, en su totalidad, se utilizaron como moldes para Safe-SeqS, tal como

se describe en la figura 3. En siete experimentos independientes de este tipo, el número de familias de UID identificadas mediante secuenciación fue de 624.678 ± 421.274, lo que es consistente con una eficiencia de amplificación del 92 ± 9,6% por ronda de PCR.

El fabricante indica que la tasa de error de la polimerasa Phusion, estimada mediante la clonación de productos de PCR que codifican la β-galactosidasa en vectores plasmídicos y la transformación en bacterias es de 4,4 x 10^{-7} errores/pb/ciclo de PCR. Un análisis convencional, incluso con una identificación de bases muy estricto de la secuenciación de Illumina reveló una tasa de error aparente de 9,1 x 10^{-6} errores/pb/ciclo de PCR, más de un orden de magnitud mayor que la tasa de error de la polimerasa de Phusion (tabla 2A). Por el contrario, la Safe-SeqS de los mismos datos reveló una tasa de error de 4,5 x 10^{-7} errores/pb/ciclo de PCR, casi idéntica a la medida para la polimerasa Phusion en ensayos biológicos (tabla 2A). La gran mayoría (>99%) de estos errores fueron sustituciones de una sola base (tabla 3A), en consonancia con los datos previos sobre el espectro de mutación creado por otras ADN polimerasas procariotas (15, 46, 47).

Tabla 2A–2C. Safe–SeqS con UID exógenos

2A. Fidelidad de la polimerasa	**Media**	**Desviación estándar**
Análisis convencional de 7 repeticiones		
Pb de alta calidad	996.855.791	64.030.757
Mutaciones totales identificadas	198.638	22.515
Mutaciones/pb	**2,0E–04**	**1,7.E–05**
Tasa de error de Phusion calculada (errores/pb/ciclo)	**9,1E–06**	**7,7E–07**
Análisis Safe-SeqS de 7 repeticiones		
Pb de alta calidad	996.855.791	64.030.757
Familias de UID	624.678	421.274
Miembros/familia de UID	107	122
Supermutantes totales identificados	197	143
Supermutantes/pb	**9,9E–06**	**2,3E–06**
Tasa de error de Phusion calculada (errores/pb/ciclo)	**4,5E–07**	**1,0E–07**
2B. Mutaciones de *CTNNB1* en el ADN de células humanas normales		
Análisis convencional de 3 individuos		
Pb de alta calidad	559.334.774	66.600.749
Mutaciones totales identificadas	118.488	11.357
Mutaciones/pb	**2,1E–04**	**1,6:E–05**
Análisis Safe-SeqS de 3 individuos		
Pb de alta calidad	559.334.774	66.600.749
Familias de UID	374.553	263.105
Miembros/familia de UID	68	38
Supermutantes totales identificados	99	78
Supermutantes/pb	**9,0E–06**	**3,1E–06**
2C. Mutaciones mitocondriales en el ADN de células humanas normales		
Análisis convencional de 7 individuos		
Pb de alta calidad	147.673.456	54.308.546
Mutaciones totales identificadas	30.599	12.970
Mutaciones/pb	**2,1E–04**	**9,4E–05**
Análisis Safe-SeqS de 7 individuos		
Pb de alta calidad	147.673.456	54.308.546
Familias de UID	515.600	89.985
Miembros/familia de UID	15	6
Supermutantes totales identificados	135	61
Supermutantes/pb	**1,4E–05**	**6,8E–06**

Tabla 3A–C. Fracción de sustituciones, inserciones y supresiones de una sola base con UID exógenos

3A. Fidelidad de la polimerasa	**Media**	**Desviación estándar**
Análisis convencional de 7 repeticiones		
Mutaciones totales identificadas	198.638	22.515
Fracción de mutaciones representadas por sustituciones de una sola base	99%	0%

3A. Fidelidad de la polimerasa	**Media**	**Desviación estándar**
Análisis convencional de 7 repeticiones		
Fracción de mutaciones representadas por deleciones	1%	0%
Fracción de mutaciones representadas por inserciones	0%	0%
Análisis Safe-SeqS de 7 repeticiones		
Supermutantes totales identificados	197	143
Fracción de supermutantes representadas por sustituciones de una sola base	99%	2%
Fracción de supermutantes representados por deleciones	1%	2%
Fracción de supermutantes representados por inserciones	0%	0%

3B. Mutaciones de *CTNNB1* en el ADN de células humanas normales		
Análisis convencional de 3 individuos		
Mutaciones totales identificadas	118.488	11.357
Fracción de mutaciones representadas por sustituciones de una sola base	97%	0%
Fracción de mutaciones representadas por deleciones	3%	0%
Fracción de mutaciones representadas por inserciones	0%	0%
Análisis Safe-SeqS de 3 individuos		
Supermutantes totales identificados	99	78
Fracción de supermutantes representadas por sustituciones de una sola base	100%	1%
Fracción de supermutantes representados por deleciones	0%	1%
Fracción de supermutantes representados por inserciones	0%	0%

3C. Mutaciones mitocondriales en el ADN de células humanas normales		
Análisis convencional de 7 individuos		
Mutaciones totales identificadas	30.599	12.970
Fracción de mutaciones representadas por sustituciones de una sola base	98%	1%
Fracción de mutaciones representadas por deleciones	2%	1%
Fracción de mutaciones representadas por inserciones	0%	0%
Análisis Safe-SeqS de 7 individuos		
Supermutantes totales identificados	135	61
Fracción de supermutantes representadas por sustituciones de una sola base	99%	1%
Fracción de supermutantes representados por deleciones	1%	1%
Fracción de supermutantes representados por inserciones	0%	0%

Safe-SeqS también permitió una determinación del número total de acontecimientos mutacionales distintos y una estimación del ciclo de PCR en el que se produjo la mutación. Hubo 19 ciclos de PCR realizada en pocillos que contenían una sola molécula molde en estos experimentos. Si se produjera un error de polimerasa en el ciclo 19, solo se produciría un supermutante (de la hebra que contenía la mutación). Si el error se produjera en el ciclo 18 debería haber dos supermutantes (derivados de las hebras mutantes producidas en el ciclo 19), etc. Por consiguiente, el ciclo en el que se produjo el error está relacionado con el número de supermutantes que contienen ese error. Los datos de siete experimentos independientes demuestran un número relativamente consistente de errores de polimerasa totales observados ($2,2 \pm 1,1 \times 10^{-6}$ mutaciones distintas/pb), de acuerdo con el número esperado de observaciones de las simulaciones ($1,5 \pm 0,21 \times 10^{-6}$ mutaciones distintas/pb). Los datos también

muestran una cronología muy variable de la aparición de errores de polimerasa entre los experimentos (tabla 4), tal como se predijo a partir del análisis clásico de la fluctuación (1). Este tipo de información es difícil de derivar utilizando el análisis convencional de los mismos datos de secuenciación de próxima generación, en parte debido a la tasa de mutación aparente prohibitivamente alta observada anteriormente.

Tabla 4A–4G. Número observado y esperado de errores generados por la polimerasa Phusion

4A. Experimento 1	**Observado**	**Esperado (media ± SD)***
Mutaciones representadas por 1 supermutante	10	19 ± 3,7
Mutaciones representadas por 2 supermutantes	8	5,8 ± 2,3
Mutaciones representadas por 3 supermutantes	4	1,3 ± 1,1
Mutaciones representadas por 4 supermutantes	4	1,8 ± 1,3
Mutaciones representadas por 5 supermutantes	2	0,61 ± 0,75
Mutaciones representadas por 6 supermutantes	2	0,22 ± 0,44
Mutaciones representadas por 7 supermutantes	0	0,01 ± 0,10
Mutaciones representadas por 8 supermutantes	0	0,87 ± 0,86
Mutaciones representadas por 9 supermutantes	2	0,28 ± 0,51
Mutaciones representadas por 10 supermutantes	0	0,14 ± 0,38
Mutaciones representadas por > 10 supermutantes	3	1,5 ± 2,7
Mutaciones distintas	35	32 ± 4,2
4B. Experimento 2		
Mutaciones representadas por 1 supermutante	19	23 ± 4,1
Mutaciones representadas por 2 supermutantes	5	9,5 ± 2,8
Mutaciones representadas por 3 supermutantes	4	2,7 ± 1,6
Mutaciones representadas por 4 supermutantes	7	2,7 ± 1,7
Mutaciones representadas por 5 supermutantes	2	0,88 ± 0,94
Mutaciones representadas por 6 supermutantes	1	0,40 ± 0,60
Mutaciones representadas por 7 supermutantes	3	0,16 ± 0,42
Mutaciones representadas por 8 supermutantes	1	0,99 ± 1,0
Mutaciones representadas por 9 supermutantes	1	0,39 ± 0,68
Mutaciones representadas por 10 supermutantes	0	0,17 ± 0,43
Mutaciones representadas por > 10 supermutantes	9	1,8 ± 3,4
Mutaciones distintas	52	43 ± 5,1
4C. Experimento 3		
Mutaciones representadas por 1 supermutante	7	17 ± 3,4
Mutaciones representadas por 2 supermutantes	9	5,4 ± 2,0
Mutaciones representadas por 3 supermutantes	4	1,2 ± 1,1
Mutaciones representadas por 4 supermutantes	4	1,7 ± 1,4
Mutaciones representadas por 5 supermutantes	2	0,50 ± 0,70
Mutaciones representadas por 6 supermutantes	0	0,17 ± 0,45
Mutaciones representadas por 7 supermutantes	1	0,03 ± 0,17
Mutaciones representadas por 8 supermutantes	0	0,59 ± 0,74
Mutaciones representadas por 9 supermutantes	0	0,24 ± 0,50
Mutaciones representadas por 10 supermutantes	1	0,07 ± 0,29
Mutaciones representadas por > 10 supermutantes	5	1,5 ± 2,6
Mutaciones distintas	33	28 ± 3,7
4D. Experimento 4		
Mutaciones representadas por 1 supermutante	7	15 ± 3,7
Mutaciones representadas por 2 supermutantes	8	4,1 ± 1,7
Mutaciones representadas por 3 supermutantes	2	0,70 ± 0,74
Mutaciones representadas por 4 supermutantes	1	1,5 ± 1,3
Mutaciones representadas por 5 supermutantes	3	0,21 ± 0,52
Mutaciones representadas por 6 supermutantes	2	0,08 ± 0,27
Mutaciones representadas por 7 supermutantes	1	0,0 ± 0,0
Mutaciones representadas por 8 supermutantes	2	0,65 ± 0,77
Mutaciones representadas por 9 supermutantes	2	0,17 ± 0,43
Mutaciones representadas por 10 supermutantes	0	0,05 ± 0,22
Mutaciones representadas por > 10 supermutantes	1	0,92 ± 2,1
Mutaciones distintas	29	23 ± 3,2

4E. Experimento 5		
Mutaciones representadas por 1 supermutante	9	23 ± 4,1
Mutaciones representadas por 2 supermutantes	6	9,5 ± 2,8
Mutaciones representadas por 3 supermutantes	5	2,7 ± 1,6
Mutaciones representadas por 4 supermutantes	3	2,7 ± 1,7
Mutaciones representadas por 5 supermutantes	6	0,88 ± 0,94
Mutaciones representadas por 6 supermutantes	2	0,40 ± 0,60
Mutaciones representadas por 7 supermutantes	1	0,16 ± 0,42
Mutaciones representadas por 8 supermutantes	2	0,99 ± 1,0
Mutaciones representadas por 9 supermutantes	2	0,39 ± 0,68
Mutaciones representadas por 10 supermutantes	3	0,17 ± 0,43
Mutaciones representadas por > 10 supermutantes	7	1,8 ± 3,4
Mutaciones distintas	46	43 ± 5,1
4F. Experimento 6		
Mutaciones representadas por 1 supermutante	4	6,7 ± 2,8
Mutaciones representadas por 2 supermutantes	7	1,5 ± 1,2
Mutaciones representadas por 3 supermutantes	1	0,10 ± 0,33
Mutaciones representadas por 4 supermutantes	2	0,60 ± 0,82
Mutaciones representadas por 5 supermutantes	0	0,07 ± 0,26
Mutaciones representadas por 6 supermutantes	0	0,01 ± 0,10
Mutaciones representadas por 7 supermutantes	1	0,0 ± 0,0
Mutaciones representadas por 8 supermutantes	1	0,39 ± 0,60
Mutaciones representadas por 9 supermutantes	0	0,01 ± 0,10
Mutaciones representadas por 10 supermutantes	0	0,0 ± 0,0
Mutaciones representadas por > 10 supermutantes	2	0,50 ± 1,1
Mutaciones distintas	18	9,9 ± 1,4
4G. Experimento 7		
Mutaciones representadas por 1 supermutante	8	2,9 ± 1,6
Mutaciones representadas por 2 supermutantes	2	0,61 ± 0,79
Mutaciones representadas por 3 supermutantes	0	0,04 ± 0,24
Mutaciones representadas por 4 supermutantes	0	0,41 ± 0,59
Mutaciones representadas por 5 supermutantes	1	0,01 ± 0,10
Mutaciones representadas por 6 supermutantes	0	0,0 ± 0,0
Mutaciones representadas por 7 supermutantes	0	0,0 ± 0,0
Mutaciones representadas por 8 supermutantes	0	0,14 ± 0,35
Mutaciones representadas por 9 supermutantes	0	0,01 ± 0,10
Mutaciones representadas por 10 supermutantes	0	0,0 ± 0,0
Mutaciones representadas por > 10 supermutantes	0	0,32 ± 0,93
Mutaciones distintas	11	4,5 ± 0,62

*Véase el texto SI para detalles de las simulaciones

EJEMPLO 4 – Análisis de la composición de oligonucleótidos.

Se puede tolerar un pequeño número de errores durante la síntesis de oligonucleótidos a partir de precursores de fosforoamidita para la mayoría de las aplicaciones, tales como la PCR rutinaria o la clonación. Sin embargo, para la biología sintética, en la que muchos oligonucleótidos deben unirse, tales errores representan un obstáculo importante para el éxito. Se han ideado estrategias inteligentes para hacer más eficiente el proceso de construcción de genes (48, 49), pero todas estas estrategias se beneficiarían de una síntesis más precisa de los propios oligonucleótidos. La determinación del número de errores en los oligonucleótidos sintetizados es difícil debido a que la fracción de oligonucleótidos que contienen errores puede ser inferior a la sensibilidad de los análisis convencionales de secuenciación de próxima generación.

Para determinar si Safe-SeqS podría utilizarse para esta determinación, se utilizó la química estándar de fosforoamidita para sintetizar un oligonucleótido que contenía 31 bases que se diseñaron para ser idénticas a las analizadas en el experimento de fidelidad de la polimerasa descrito anteriormente. En el oligonucleótido sintético, las 31 bases estaban rodeadas por secuencias complementarias a los cebadores que podrían utilizarse para las etapas de asignación de UID de Safe-SeqS (figura 3). Al realizar Safe-SeqS en ~300.000 oligonucleótidos, los presentes inventores descubrieron que había $8,9 \pm 0,28 \times 10^{-4}$ supermutantes/pb y que estos errores se produjeron a lo largo

de los oligonucleótidos (figura 6A). Los oligonucleótidos contenían un gran número de errores de inserción y deleción, representando el 8,2 ± 0,63% y el 25 ± 1,5% de los supermutantes totales, respectivamente. Es importante destacar que tanto la posición como la naturaleza de los errores eran altamente reproducibles entre siete replicaciones independientes de este experimento realizado en el mismo lote de oligonucleótidos (figura 6A). Esta naturaleza y distribución de errores tenían muy poco en común con la de los errores producidos por la polimerasa Phusion (figura 6B y tabla 5), que se distribuyeron en el patrón estocástico esperado entre los experimentos repetidos. El número de errores en los oligonucleótidos sintetizados con fosforoamiditas fue -60 veces mayor que en los productos equivalentes sintetizados por la polimerasa Phusion. Estos datos, *in toto,* indican que la gran mayoría de los errores en los primeros se generaron durante su síntesis en lugar de durante el procedimiento Safe-SeqS.

Tabla 5. ADN sintetizado con fosforoamidita frente a Phusion: Comparación de transiciones frente a transversiones

Fosforoamiditas	**Exp. 1**	**Exp.2**	**Exp.3**	**Exp.4**	**Exp.5**	**Exp.6**	**Exp. 7**	**Promedio**	**Desviación estándar**
Supermutantes de transición	496	509	471	396	323	273	470	**420**	**92**
Supermutantes de transversión	1.494	1.499	1.521	1.154	944	907	1.626	**1.306**	**298**
Valor p*	**3,4E–05**								
Phusion									
Supermutantes de transición	63	275	127	5	87	182	103	**120**	**87**
Supermutantes de transversión:	14	124	77	12	57	191	63	**77**	**63**
Valor p*	**0,08**								

*Los valores de p se calcularon utilizando una prueba t pareada de dos colas

ES 2 625 288 T3

¿Preserva Safe-SeqS la relación de secuencias mutantes:normales en los moldes originales? Para resolver esta cuestión, los presentes inventores sintetizaron dos oligonucleótidos de 31 bases de secuencia idéntica con la excepción del nt 15 (50:50 C/G en lugar de T) y se mezclaron en fracciones nominales mutantes/normales del 3,3% y el 0,33%. Mediante el análisis Safe-SeqS de las mezclas de oligonucleótidos, los presentes inventores descubrieron que las relaciones eran del 2,8% y el 0,27%, respectivamente. Los presentes inventores concluyeron que los procedimientos de asignación y amplificación de UID utilizados en Safe-SeqS no alteran en gran medida la proporción de secuencias de variantes y, por lo tanto, proporcionan una estimación fiable de esa proporción cuando se desconoce. Esta conclusión también es apoyada por la reproducibilidad de las fracciones variantes cuando se analizan en experimentos independientes de Safe-SeqS (figura 6A).

EJEMPLO 5 – Análisis de secuencias de ADN de células humanas normales

La estrategia de UID exógeno (figura 3) se utilizó a continuación para determinar la prevalencia de las mutaciones raras en una pequeña región del gen *CTNNBI* de ~100.000 células humanas normales de tres individuos no relacionados. A través de la comparación con el número de familias de UID obtenido en los experimentos de Safe-SeqS (tabla 2B), los presentes inventores calcularon que la mayoría (78 ± 9,8%) de los fragmentos de entrada se convirtieron en familias de UID. Había un promedio de 68 miembros/familia de UID que cumplían fácilmente la redundancia requerida para Safe-SeqS (figura 7). El análisis convencional de los datos de secuenciación de Illumina reveló un promedio de 118.488 ± 11.357 mutaciones entre los ~560 Mb de secuencia analizada por muestra, lo que corresponde a una prevalencia aparente de mutación de $2,1 \pm 0,16 \times 10^{-4}$ mutaciones/pb (tabla 2B). Solo se observó un promedio de 99 ± 78 supermutantes en el análisis Safe-SeqS. La gran mayoría (> 99%) de supermutantes fueron sustituciones de una sola base y la tasa de mutación calculada fue de $9,0 \pm 3,1 \times 10^{-6}$ mutaciones/pb (tabla 3B). Por tanto, Safe-SeqS redujo la frecuencia aparente de mutaciones en el ADN genómico, como mínimo, 24 veces (figura 4).

Una posible estrategia para aumentar la especificidad de Safe-SeqS es realizar la amplificación de la biblioteca (y, posiblemente, los ciclos de asignación de UID) en múltiples pocillos. Esto puede lograrse en tan solo 2 o hasta 384 pocillos utilizando placas de PCR estándar o ampliarse hasta muchos más pocillos cuando se usa un dispositivo microfluídico (de miles a millones). Cuando se realiza de esta manera, se pueden introducir secuencias de indexaxión en los moldes que son únicos para los pocillos en los que se amplifica el molde. Por tanto, las mutaciones raras deberían dar lugar a dos supermutantes (es decir, uno de cada hebra), ambos con la misma secuencia de índice de pocillos. Al realizar Safe-SeqS con UID exógenos en los moldes de *CTNNBI* descritos anteriormente y diluidos en 10 pocillos (cada pocillo produce moldes amplificados con una secuencia de índice diferente), la tasa de mutación se redujo adicionalmente de $9,0 \pm 3,1 \times 10^{-6}$ a $3,7 \pm 1,2 \times 10^{-6}$ supermutantes/pb. Por lo tanto, el análisis de los moldes en múltiples compartimentos, de una manera que produce moldes codificados de forma diferente basados en el compartimiento en el que se amplificaron moldes, puede ser una estrategia adicional para aumentar la especificidad de Safe-SeqS.

EJEMPLO 6 – Análisis de secuencias de ADN de ADN mitocondrial

Los presentes inventores aplicaron la estrategia idéntica a un segmento corto de ADN mitocondrial en ~1.000 células de cada uno de siete individuos no relacionados. El análisis convencional de las bibliotecas de secuenciación de Illumina producidas con el procedimiento Safe–SeqS (figura 3) reveló un promedio de 30.599 ± 12.970 mutaciones entre los ~150 Mb de secuencia analizada por muestra, lo que corresponde a una prevalencia aparente de mutación de $2,1 \pm 0,94 \times 10^{-4}$ mutaciones/pb (tabla 2C). Solo se observaron 135 ± 61 supermutantes en el análisis Safe-SeqS. Como con el gen *CTNNB1*, la gran mayoría de las mutaciones fueron sustituciones de una sola base, aunque también se observaron deleciones ocasionales de una sola base (tabla 3C). La tasa de mutación calculada en el segmento analizado de ADNmt fue de $1,4 \pm 0,68 \times 10^{-5}$ mutaciones/pb (tabla 2C). Por tanto, Safe-SeqS redujo la frecuencia aparente de mutaciones en el ADN genómico, como mínimo, 15 veces.

EJEMPLO 7 – Materiales y procedimientos

UID endógenos. El ADN genómico del páncreas humano o de células linfoblastoides cultivadas se preparó utilizando kits Qiagen. Se utilizó ADN de páncreas para el experimento de captura y se utilizaron las células linfoblastoides para el experimento de PCR inversa. El ADN se cuantificó mediante absorbancia óptica y con PCRc. El ADN se fragmentó a un tamaño promedio de ~200 pb mediante cizallamiento acústico (Covaris), después se repararon los extremos, se introdujeron colas de A y se ligaron a adaptadores en forma de Y de acuerdo con los protocolos estándar de Illumina. Los extremos de cada molécula molde proporcionan UID endógenos correspondientes a sus posiciones cromosómicas. Después de la amplificación mediada por PCR de las bibliotecas con secuencias de cebadores dentro de los adaptadores, se capturó el ADN (1) con un filtro que contenía 2.594 nt correspondiente a seis genes de cáncer. Después de la captura, se realizaron 18 ciclos de PCR para asegurar cantidades suficientes de molde para la secuenciación en un instrumento Illumina GA IIx.

Para los experimentos de PCR inversa (figura 5), los presentes inventores ligaron adaptadores personalizados (IDT, tabla 6) en lugar de adaptadores estándar de Illumina en forma de Y al ADN celular cortado. Estos adaptadores conservaban la región complementaria del cebador de secuenciación universal pero carecían de las secuencias de

injerto requeridas para la hibridación con la célula de flujo Illumina GA IIx. El ADN ligado se diluyó en 96 pocillos y el ADN en cada columna de 8 pocillos se amplificó con un cebador directo único que contenía una de 12 secuencias índice en su extremo 5' más un cebador inverso estándar (tabla 6). Las amplificaciones se realizaron utilizando Phusion HotStart Í (NEB) en 50 µl de reacciones que contenían tampón Phusion HF 1X, dNTP 0,5 mM, 0,5 µM de cada cebador directo e inverso (ambos fosforilados en 5') y 1U de la polimerasa Phusion. Se utilizaron las siguientes condiciones de ciclado: un ciclo de 98ºC durante 30 segundos; y 16 ciclos de 98ºC durante 10 segundos, 65ºC durante 30 segundos y 72ºC durante 30 segundos. Las 96 reacciones se agruparon y, después, se purificaron utilizando un kit de purificación de PCR Qiagen MinElute (No. de cat. 28004) y un kit QIAquick Gel Extraction (No. De cat. 28704). Para preparar los moldes circulares necesarios para la PCR inversa, el ADN se diluyó hasta ~1 ng/ul y se ligó con ADN ligasa de T4 (Enzymatics) durante 30 minutos a temperatura ambiente en una reacción de 600 µl que contenía 1X de tampón de unión de ADN de T4 y 18.000 U de ADN ligasa de T4. La reacción de ligadura se purificó utilizando un kit Qiagen MinElute. La PCR inversa se realizó utilizando Phusion Hot Start I en 90 ng de molde circular distribuido en doce reacciones de 50 µl, conteniendo cada una 1X de tampón Phusion HF, dNTP 0,25 mM, 0,5 µM de cada uno de los cebadores *KRAS* directos e inversos (tabla 6) y 1 U de polimerasa Phusion. Los cebadores específicos de *KRAS* contenían ambos secuencias de injerto para la hibridación a la célula de flujo Illumina GA IIx (tabla 6). Se utilizaron las siguientes condiciones de ciclado: un ciclo de 98ºC durante 2 minutos; y 37 ciclos de 98ºC durante 10 segundos, 61ºC durante 15 segundos y 72ºC durante 10 segundos. La purificación final se realizó con un kit NucleoSpin Extract II (Macherey-Nagel) y se eluyó en 20 ul de tampón NE. Los fragmentos de ADN resultantes contenían UID compuestos por tres secuencias: dos endógenas, representadas por los dos extremos de los fragmentos cortados originales más la secuencia exógena introducida durante la amplificación de indexación. Dado que se utilizaron 12 secuencias exógenas, esto aumentó el número de UID distintos 12 veces más que lo que se obtuvo sin UID exógenos. Este número podría aumentarse fácilmente utilizando un mayor número de cebadores distintos.

Tabla 6. Oligonucleótidos utilizados

Leyenda de la fuente:
REGÍON COMPLEMENTARIA A LOS MOLDES
SECUENCIA DE UID ESPECÍFICA DEL MOLDE
SECUENCIA UNIVERSAL
SECUENCIA ÍNDICE ESPECÍFICA DEL EXPERIMENTO

Leyenda del símbolo:
/5Fos/ = 5' Fosfato
* = Enlace fosforotioato

CEBADORES DE INJERTO DE ILLUMINA (PARA HIBRIDACIÓN CON LA CÉLULA DE FLUJO)

UID endógenos

Captura	**Secuencia (SEQ ID NO:1–81, respectivamente)**
Adaptador - hebra 1	/5Fos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adaptador - hebra 2	ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo - dir	AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCT TCCGAT*C*T
Amplificación del genoma completo - inv	CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGA T*C*T
Amplificación posterior a la captura - dir	AATGATACGGCGACCACCGAGATCTACACACTCTTTCCCTACACGACGCTCT TCCGAT*C*T
Amplificación posterior a la captura - inv	CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGA T*C*T
Cebador de secuenciación, lectura 1 (Illumina; San Diego, CA)	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Cebador de secuenciación, lectura 2 (Illumina; San Diego, CA)	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

PCR inversa	
Adaptador - hebra 1	/5Fos/GATCGGAAGAGCGGITCAGCAGGAATGCCGAG
Adaptador - hebra 2	ACACTCTTTCCCTACACGAOGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-1	/5Fos/**CGTGAT**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-2	/5Fos/**ACATCG**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-3	/5FosB/**GCCTAA**ACACTCriTCCCrACACGACGCTCrTCCGAT*C*T
Amplificación del genoma completo – dir-4	/5Fos/**TGGTCA**ACACTCTTTCCCTACACGACGCTCTTCCGAr*C*T
Amplificación del genoma completo – dir-5	/5Fos/**CACTGT**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-6	/5Fos/**ATTGGC**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-7	/5Fos/**GATCTG**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-8	/5Fos/**TCAAGT**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-9	/5Fos/**CTGATC**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-10	/5Fos/**AAGCTA**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-11	/5Fos/**GTAGCC**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo – dir-12	/5Fos/**TACAAG**ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Amplificación del genoma completo - inv	/5Fos/CTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T
PCR inversa - antisentido	**AATGATACGGCGACCACCCGAGATCTACAC***CAGCAGGCCTTATAATAAAAATAATGA*
PCR inversa – dir	**CAAGCAGAAGACGGCATACGAGAT***TGACTGAATATAAACTTGTGGTAGTTG*

PCR inversa	
Cebador de secuenciación 1 (para leer secuencias internas)	ACACTCTTTCCCTACACGACGCTCTICCGATCT
Cebador de secuenciación 2 (para leer secuencias internas)	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Cebador índice 1 (para leer índices experimentales)	CGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Cebador índice 2 (para leer índices experimentales)	CGGAAGAGCGGTTCAGCAGGAATGCCGAG

UID exógenos

Fidelidad de la polimerasa	
Amplificación por PCR digital - dir	*GGTTACAGGCTCATGATGTAACC*
Amplificación PCR digital – inv	*GATACCAGCTTGGTAATGGCA*
Amplificación para asignación de UID - dir	CGACGTAAAACGACGGCCAGT***NNNNNNNNNNNN****GGTTACAGGCTCATGATGTAACC*
Amplificación para asignación de UID - inv	CACACAGGAAACAGCTATGACCATG*GATACCAGCTTGGTAATGGCA*
Amplificación de la biblioteca – dir-1	AATGATACGGCGACCACCGAGATCTACACGCTGACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-2	AATGATACGGCGACCACCGAGATCTACACATCGCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-3	AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-4	AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-5	AATGATACGGCGACCACCGAGATCTACACGACCGTCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-6	AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-7	AATGATACGGCGACCACCGAGATCTACACGATCTGCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-8	AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-9	AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-10	AATGATACGGCGACCACCGAGATCTACACAAGCTACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – inv	**CAAGCAGAAGACGGCATACGAGAT**CACACAGGAAACAGCTATGACCA*T*G
Cebador de secuenciación (para leer UID y secuencias internas)	CGACGTAAAACGACGGCCAGT
Cebador índice (para leer índices experimentales)	ACTGGCCGTCGTTTTACGTCG

Mutaciones de *CTNNBI* en el ADN de células humanas normales	
Amplificación para asignación de UID - dir	CGACGTAAAACGACGGCCAGT*NNNNNNNNNNNNNNNNGCAGCLACAGTCTTACCTGGA CT*
Amplificación para asignación de UID - inv	CACACAGGAAACAGCTATGACCATGTCCACATCCTCTTCCTCAGGATT
Amplificación de la biblioteca – dir	AATGATACGGCGACCACCGAGATCTACACCGACGTAAAACGACGGCCA*C*T
Amplificación de la biblioteca – inv-1	CAAGCAGAAGACGGCATACGAGATATCACGCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-2	CAAGCAGAAGACGGCATACGAGATCGATGTCACACAGGAAACAGCTATGACCA* T*G

UID exógenos

Fidelidad de la polimerasa

Amplificación de la biblioteca – inv-3	CAAGCAGAAGACGGCATACGAGATTGACCACACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-4	CAAGCAGAAGACGGCATACGAGATGCAATCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-5	CAAGCAGAAGACGGCATACGAGATCAGATCCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-6	CAAGCAGAAGACGGCATACGAGATACTTGACACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-7	CAAGCAGAAGACGGCATACGAGATGACCAGCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-8	CAAGCAGAAGACGGCATACGAGATTAGCTCCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-9	CAAGCAGAAGACGGCATACGAGATGCCTACCACACAGGAAACAGCTATGACCA* T*G
Amplificación de la biblioteca – inv-10	CAAGCAGAAGACGGCATACGAGATCTCGTACACACAGGAAACAGCTATGACCA* T*G
Cebador de secuenciación (para leer UID y secuencias internas)	CGACGTAAAACGACGGCCAGT
Cebador índice (para leer índices experimentales)	CATGGTCATAGCTGTTTCCTGTGTG

Mutaciones mitocondriales en el ADN de células humanas normales

Amplificación para asignación de UID - dir	CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNNNNTACCGAGAAGCTCICAAGA A
Amplificación para asignación de UID - inv	CACACAGGAAACAGCTATGACCATG*ATGCTAAGGCGAGGATGAAA*
Amplificación de la biblioteca – dir-1	AATGATACGGCGACCACCGAGATCTACACACATCGCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-2	AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-3	AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-4	AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-5	AATGATACGGCGACCACCGAGATCTACACGATCTCCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-6	AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – dir-7	AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCC A*G*T
Amplificación de la biblioteca – inv	**CAAGCAGAAGACGGCATACGAGAT**CACACAGGAAACAGCTATGACCA*T*G
Cebador de secuenciación 1 (para leer UID)	CGACGTAAAACGACGGCCAGT
Cebador de secuenciación 2 (para leer secuencias internas)	CCTAATTCCCCCCATOCTTAC
Cebador índice (para leer índices experimentales)	ACTGGCCGTCGTTTTACGTCG

Análisis de la composición de oligonucleótidos de fosforoamidita	
Molde sintetizado, sv	*GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTGTAACTTTAAAACATATTTTGCCA TTACCAAGCTGGTATC*
Molde sintetizado, mut (S = 50/50 mezcla de C y G)	*GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTGSAACTTTAAAACATATTTTGCCA TTACCAAGCTGGTATC*
Amplificación para asignación de UID - dir	ACACTCTTTCCCTACACGACGCTC***NNNNNNNNNNNN****GGTGAGTCTGTGCAGGCAT*
Amplificación para asignación de UID - inv	*CTCGAGCACTGTCCTGACTGAGACGATACCAGCTTGGTAATGGCA*
Amplificación de la biblioteca – dir	AATGATACGGCGACCACCGAGATCTACACCGTGATACACTCTTTCCCTACACGA CGCTC*T*C
Amplificación de la biblioteca – inv	**CAAGCAGAAGACGGCATACGAGAT**CTCGAGCACTGTCCTGACTGAG*A*C
Cebador de secuenciación (para leer UID y secuencias internas)	ACACTCTTTCCCTACACGACGCTC

ES 2 625 288 T3

UID exógenos Se preparó ADN genómico a partir de mucosas colónicas normales humanas o linfocitos de sangre utilizando kits Qiagen. El ADN de la mucosa colónica se utilizó para los experimentos con *CTNNB1* y ADN mitocondrial, mientras que el ADN linfocitario se utilizó para los experimentos en *CTNNB1* y sobre la fidelidad de la polimerasa. El ADN se cuantificó con PCR digital (2) utilizando cebadores que amplificaron los genes de una sola copia de células humanas (análisis de la fidelidad de la polimerasa y *CTNNB1*), PCRc (ADN mitocondrial) o mediante absorbancia óptica (oligonucleótidos). Cada cadena de cada molécula molde se codificó con un UID de 12 ó 14 bases utilizando dos ciclos de PCR específica de amplicón, tal como se describe en el texto y en la figura 3. Los cebadores específicos de amplicón contenían ambos secuencias marcadoras universales en sus extremos 5' para una etapa de amplificación posterior. Los UID constituyeron 12 ó 14 secuencias de nucleótidos al azar añadidas al extremo 5' de los cebadores directos específicos de amplicón (tabla 6). Estos cebadores pueden generar 16,8 y 268 millones de UID distintos, respectivamente. Es importante que el número de UID distintos exceda en gran medida el número de moléculas molde originales para minimizar la probabilidad de que dos moldes originales diferentes adquirieran el mismo UID. Los ciclos de PCR de asignación de UID incluyeron Phusion Hot Start III (NEB) en una reacción de 45 µl que contenía 1X tampón Phusion HF, dNTP 0,25 mM, 0,5 µM de cada cebador directo (que contenía 12-14 Ns) e inversos y 2U de polimerasa Phusion. Para mantener las concentraciones finales del molde <1,5 ng/µl, se utilizaron múltiples pocillos para crear algunas bibliotecas. Se utilizaron las siguientes condiciones de ciclado: una incubación de 98ºC durante 30 segundos (para activar la Phusion Hot Start II); y dos ciclos de 98ºC durante 10 segundos, 61ºC durante 120 segundos y 72ºC durante 10 segundos. Para asegurar la eliminación completa de los cebadores de la primera ronda, cada pocillo se digirió con 60 U de una nucleasa específica de ADN de una sola hebra (exonucleasa-I; Enzymatics) a 37ºC durante 1 hora. Después de una inactivación con calor de 5 minutos a 98ºC, se añadieron cebadores complementarios a los marcadores universales introducidos (tabla 6) a una concentración final de 0,5 µM cada uno. Estos cebadores contenían dos fosforotioatos terminales para hacerlos resistentes a cualquier actividad residual de Exonucleasa-I. También contenían secuencias de injerto en 5' necesarias para la hibridación a la célula de flujo Illumina GA IIx. Finalmente, contenían una secuencia índice entre la secuencia de injerto y la secuencia marcadora universal. Esta secuencia de índice permite analizar simultáneamente los productos de PCR de múltiples individuos diferentes en el mismo compartimento de la célula de flujo del secuenciador. Se utilizaron las siguientes condiciones de ciclado para los siguientes 25 ciclos de PCR: 98ºC durante 10 segundos y 72ºC durante 15 segundos. No se realizaron etapas de purificación intermedias en un esfuerzo para reducir las pérdidas de moléculas molde.

Después de la segunda ronda de amplificación, los pocillos se consolidaron y purificaron utilizando un kit de purificación por PCR Qiagen QIAquick (No. de cat. 28104) y se eluyeron en 50 µl de tampón EB (Qiagen). Los fragmentos del tamaño esperado se purificaron después de la electroforesis en gel de agarosa (bibliotecas de ADNmt) o de poliacrilamida (todas las demás bibliotecas). Para la purificación en gel de agarosa, las ocho alícuotas de 6 µl se cargaron en pocillos de un Gel Size Select al 2% (Invitrogen) y se recogieron bandas del tamaño esperado en tampón EB según lo especificado por el fabricante. Para la purificación en gel de poliacrilamida, se cargaron diez alícuotas de 5 µl en pocillos de un gel de poliacrilamida TBE al 10% (Ínvitrogen). Las láminas de gel que contenían los fragmentos de interés se escindieron, se trituraron y se eluyeron esencialmente, tal como se ha descrito (3).

Análisis de la fidelidad de la polimerasa Phusion. La amplificación de un fragmento de ADN genómico humano dentro del gen BMX (RefSeq de acceso NM_203281.2) se realizó primero utilizando las condiciones de PCR descritas anteriormente. El molde se diluyó de manera que un promedio de una molécula molde estaba presente en cada uno de 10 pocillos de una placa de PCR de 96 pocillos. A continuación se realizaron reacciones de cincuenta ul en 1X de tampón Phusion HF, dNTP 0,25 mM, 0,5 µM de cada uno de los cebadores directos e inversos (tabla 6) y 2U de la polimerasa Phusion. Las condiciones de ciclado fueron un ciclo de 98ºC durante 30 segundos; y 19 ciclos de 98ºC durante 10 segundos, 61ºC durante 120 segundos y 72ºC durante 10 segundos. Los cebadores se retiraron mediante digestión con 60 U de Exonucleasa-I a 37ºC durante 1 hora, seguido de una inactivación con calor de 5 minutos a 98ºC. No se realizó ninguna purificación del producto de PCR, ya sea antes o después de la digestión con Exonucleasa-I. El contenido completo de cada pocillo se utilizó después como moldes para la estrategia de UID exógenos descritos anteriormente.

Secuenciación. La secuenciación de todas las bibliotecas descritas anteriormente se realizó utilizando un instrumento Illumina GA IIx según lo especificado por el fabricante. La longitud total de las lecturas utilizadas para cada experimento varió de 36 a 73 bases. Se realizó identificación de bases y alineación de secuencias con la gama Eland (Illumina). Solo se utilizaron lecturas de alta calidad que cumplían los siguientes criterios para el análisis posterior: (i) las primeras 25 bases pasaron el filtro de castidad estándar de Illumina; (ii) cada base en la lectura tenía una puntuación de calidad ≥20; y (iii) ≤ 3 apareamientos erróneos con las secuencias esperadas. Para las bibliotecas de UID exógenos, los presentes inventores requirieron, además, que los UID tuvieran una puntuación de calidad ≥30. Los presentes inventores observaron una frecuencia relativamente alta de errores en los extremos de las lecturas en las bibliotecas de UID endógenos preparados con el protocolo estándar de Illumina, presumiblemente introducidos durante el cizallamiento o la reparación de extremos, de modo que las primeras y últimas tres bases de estos marcadores se excluyeron del análisis.

Análisis Safe–SeqS Las lecturas de alta calidad se agruparon en familias de UID basadas en sus UID endógenos o exógenos. Solo se consideraron las familias de UID con dos o más miembros. Tales familias de UID incluyeron la gran mayoría (≥ 99%) de las lecturas de secuenciación. Para asegurar que se utilizaron los mismos datos para el

análisis convencional y Safe–SeqS, los presentes inventores también excluyeron las familias de UID que contenían solo un miembro del análisis convencional. Además, solo identificaron una base como "mutante" en el análisis de secuenciación convencional si la misma variante se identificó en, como mínimo, dos miembros de, como mínimo, una familia de UID (es decir, dos mutaciones) al comparar el análisis convencional con el de Safe-SeqS con UID exógenos. Para la comparación con Safe-SeqS con UID endógenos, los presentes inventores requirieron, como mínimo, dos miembros de cada una de dos familias de UID (es decir, cuatro mutaciones) para identificar una posición como "mutante" en el análisis convencional. Con los UID endógenos o exógenos, se definió un supermutante como una familia de UID en la que ≥ 95% de los miembros compartían la mutación idéntica. Por lo tanto, las familias de UID con <20 miembros tenía que ser 100% idénticas en la posición mutante, mientras que se permitió el 5% de replicación y tasa de error de secuenciación combinadas en las familias de UID con más miembros. Para determinar la fidelidad de la polimerasa utilizando Safe-SeqS y para comparar los resultados con análisis anteriores de la fidelidad de la polimerasa Phusion, fue necesario darse cuenta de que los análisis previos solo detectarían mutaciones presentes en ambas hebras de los productos de PCR (4). Esto sería equivalente a analizar productos de PCR generados con un ciclo menos con Safe-SeqS y la corrección apropiada se realizó en la tabla 2A. A menos que se especifique lo contrario, todos los valores enumerados en el texto y las tablas representan las medias y las desviaciones estándar.

EJEMPLO 8 – Procesos generadores de errores

Las mutaciones aparentes, definidas como cualquier identificación de base que varía con respecto a la base esperada en una posición definida, pueden ser el resultado de diversos procesos:

1. Mutaciones presentes en el ADN molde. Para los moldes derivadas de células humanas normales, estas incluyen mutaciones que estaban presentes en el cigoto, se produjeron más tarde durante el desarrollo embrionario y de los adultos, o estaban presentes en un contaminante introducido inadvertidamente en la muestra. Se espera que estas mutaciones estén presentes en ambas hebras de los moldes relevantes. Si la mutación se produjo solo en el último ciclo celular de una célula cuyo ADN se usó como molde, la mutación estaría presente en solo una hebra del molde.

2. Bases químicamente modificadas presentes en los moldes. Se ha estimado que hay muchas miles de bases oxidadas presentes en cada célula humana (5). Cuando dicho ADN se amplifica mediante la polimerasa Phusion, puede producirse una mutación aparente en una hebra.

3. Errores introducidos durante el proceso de corte requerido para generar fragmentos pequeños para la secuenciación. El cizallamiento acústico genera temperaturas elevadas de corta duración que pueden dañar el ADN.

4. Errores introducidos durante la reparación de extremos de los fragmentos cortados. La fuente de estos errores puede ser la infidelidad de la polimerasa o mediante la incorporación de bases químicamente modificadas en los dNTP utilizados para la polimerización.

5. Errores introducidos por otras etapas enzimáticas, particularmente si las enzimas son impuras y están contaminadas con nucleasas, polimerasas o ligasas.

6. Errores introducidos durante la amplificación por PCR para preparar las bibliotecas para la captura o la PCR inversa.

7. Errores durante la PCR después de la captura o durante la amplificación por PCR inversa.

8. Errores introducidos en los ciclos de asignación de UID de Safe-SeqS (figura 3).

9. Errores introducidos en los ciclos de amplificación de la biblioteca de Safe-SeqS realizados con UID exógenos. Obsérvese que si los cebadores de asignación de UID del proceso No. 8 no se eliminan completamente, podrían, potencialmente, amplificar fragmentos de ADN que contienen errores introducidos durante estos ciclos, creando un nuevo supermutante.

10. Errores introducidos en el primer ciclo de puente-PCR en la célula de flujo de Illumina. Si la amplificación es ineficiente, un error introducido en el segundo ciclo de puente-PCR también podría dar como resultado un grupo que contiene una mutación en la mayoría de sus moléculas componentes.

11. Errores en la identificación de bases.

EJEMPLO 9 – Consecución de precisión con Safe-SeqS

Con los enfoques convencionales de secuenciación por síntesis, todos los procesos de producción de errores descritos anteriormente son relevantes, lo que da como resultado un número relativamente elevado de identificaciones de mutaciones falsos positivos (tablas 1 y 2). Safe–SeqS minimiza el número de identificaciones de mutaciones falsos positivos de varias maneras. Safe–SeqS con UID exógenos da como resultado el menor número

de identificaciones de mutaciones falsos positivos porque requiere el menor número de etapas enzimáticas. Con UID exógenos, los procesos generadores de errores No. 3 a No. 7 se eliminan por completo porque estas etapas no se realizan. Safe–SeqS con UID exógenos también reduce drásticamente los errores resultantes de los procesos generadores de errores No. 10 y No. 11 debido a la forma en que se analizan los datos.

Después de Safe-SeqS con UID exógenos, los únicos errores falsos positivos que quedan deben ser los introducidos durante los ciclos de PCR de asignación de UID (proceso generador de errores No. 8) o cebadores que contienen UID residuales durante los ciclos de amplificación de la biblioteca (proceso generador de errores No. 9). Los errores del proceso generador de errores No. 8 pueden eliminarse teóricamente al requerir, como mínimo, dos supermutantes para identificar una posición como "mutante." Este requisito es razonable porque cada mutación preexistente en un molde de ADN de doble hebra debería dar lugar a dos supermutantes, uno de cada hebra. Además, este requisito eliminaría el proceso generador de errores No. 2 (bases dañadas en los moldes originales) porque dichas bases, cuando se copian, deberían dar lugar a solo un supermutante. Finalmente, los errores generados durante los ciclos de amplificación de la biblioteca (proceso No. 9) no se amplificarán con los cebadores que contienen UID residuales si dichos cebadores están completamente eliminados, tal como se realiza en el presente documento con exonucleasa-I en exceso.

Con los UID endógenos, los errores introducidos mediante los procesos No. 10 y No. 11 se reducen drásticamente debido a la forma en que se analizan los datos (como con los UID exógenos). Los errores introducidos en los procesos No. 2 a No. 7 pueden minimizarse requiriendo que se observe una mutación en, como mínimo, dos familias de UID, por las razones expuestas en el párrafo anterior. Con este requisito, en teoría se deberían identificar pocas mutaciones falsos positivos.

En la práctica, la situación se complica por el hecho de que las diversas amplificaciones no son perfectas, por lo que cada hebra de cada molécula molde original no se recupera como una familia de UID. Esta eficiencia puede variar de una muestra a otra, dependiendo en parte de la concentración de inhibidores presentes en muestras clínicas. Además, con UID exógenos, un error de polimerasa durante la etapa de amplificación de biblioteca puede crear una nueva familia de UID que no estaba representada en la etapa de asignación de UID. Si este error ocurrió en un molde mutante, se crearía un supermutante artificial adicional.

Estos factores pueden ser manipularse incorporando varios criterios adicionales en los análisis. Por ejemplo, se podría requerir que las familias de UID contengan más de dos, cinco o diez miembros. Otro requisito podría ser que los UID exógenos de supermutantes no estén relacionados con ningún otro UID en la biblioteca por una diferencia de una base. Esto eliminaría los supermutantes artificiales generados durante las etapas de amplificación de la biblioteca (indicadas en el párrafo anterior). Rutinariamente, los presentes inventores instituyeron este requisito en sus análisis Safe-SeqS, pero no implicó gran diferencia (<1%) en el número de supermutantes identificados. La especificidad para mutaciones puede incrementarse adicionalmente requiriendo más de un supermutante para identificar una posición como "mutante", tal como se ha descrito anteriormente para los UID endógenos. Cuando se requieren múltiples supermutantes, la especificidad puede aumentarse aún más exigiendo que cada cadena del molde de doble hebra original contenga la mutación o cuando las bibliotecas se amplifiquen utilizando múltiples pocillos, que las mutaciones raras comparten una secuencia introducida que identifica el pocillo en el que se amplificaron las mutaciones (es decir, una de cada hebra). Tales decisiones implican el intercambio habitual entre especificidad y sensibilidad. En los experimentos de los presentes inventores con UID exógenos (tabla 2], se requirió solo un supermutante para identificar una posición como "mutante" e incluía todas las familias de UID con más de un miembro. Como los UID endógenos se asociaron con más procesos generadores de errores que con los UID exógenos, los presentes inventores requirieron dos supermutantes para identificar una posición como mutante en los experimentos indicados en la tabla 1 y también incluyó a todas las familias de UID con más de un miembro.

EJEMPLO 10 – Prevalencias de mutación en tejidos humanos normales

Los experimentos indicados en las tablas 1 y 2, en los que se evaluaron >10.000 moldes, muestran que las mutaciones están presentes en el ADN nuclear de células humanas normales a una frecuencia de 3,5 x 10^{-6} a 9,0 x 10^{-6} mutantes/pb dependiendo de la región analizada. Es imposible determinar si este nivel bajo representa mutaciones genuinas presentes en los moldes originales o la suma de mutaciones genuinas más mutaciones artifactuales de los procesos generadores de errores descritos anteriormente. Las prevalencias de mutación en las células humanas no se han investigado ampliamente, en parte porque son muy infrecuentes. Sin embargo, se han ideado varias técnicas inteligentes para identificar mutantes raros y, en principio, pueden utilizarse para comparación. Por desgracia, las estimaciones de las prevalencias de mutaciones humanas varían ampliamente, desde incluso 10^{-5} mutantes/pb hasta un incluso 10^{-8} mutantes/pb (6-12). En varios de estos estudios, las estimaciones se complican por la falta de datos sobre la naturaleza de las mutaciones reales, podrían, en algunos casos, deberse a pérdidas de cromosomas enteros, en otros por mutaciones de sentido erróneo y en otros, principalmente, por mutaciones sin sentido o pequeñas inserciones o deleciones. Además, estos estudios utilizaron diversas fuentes de células normales y examinaron diferentes genes, lo que dificulta las comparaciones directas. Las estimaciones de las prevalencias y las tasas de mutaciones en el ADN mitocondrial varían de forma similar (13-19). Será de interés en futuros trabajos analizar los mismos moldes de ADN y genes con diversas tecnologías para determinar la base de estas diferentes estimaciones.

Pero supóngase que todas las mutaciones identificadas con Safe-SeqS representan mutaciones genuinas presentes en los moldes de ADN originales de células normales. ¿Qué dice esto acerca del número de generaciones a través de las cuales estas células han procedido desde que se concibió el organismo? Existe una relación simple entre la tasa de mutación y la prevalencia de mutación: la prevalencia de la mutación es igual al producto de la tasa de mutación y el número de generaciones que la célula ha sufrido desde la concepción. En estudios previos se ha determinado que la tasa de mutaciones somáticas es ~10^{-9} mutantes/pb/generación, aunque esta estimación también varía de un estudio a otro por razones relacionadas con las mencionadas anteriormente con respecto a la prevalencia de la mutación. La combinación de esta estimación derivada de la literatura de la tasa de mutación con las estimaciones de los presentes inventores de la prevalencia de mutación sugiere que las células normales analizadas (linfocitos, líneas celulares linfoblastoides o mucosas colónicas) se han desarrollado a través de 3.500 a 8.900 generaciones, representando las células que se dividen cada 3 a 7 días para los individuos analizados en este estudio (edad promedio de 65 años).

EJEMPLO 11 – Simulación por ordenador de errores introducidos por la polimerasa

El momento de las mutaciones introducidas por las polimerasas altera en gran medida el número final de mutaciones observadas (20). Por ejemplo, dos mutaciones diferirían en la prevalencia ~64 veces si se introducen separadas por 6 ciclos (2^6). Debido a que las polimerasas introducen mutaciones de una manera estocástica, se utilizó un simple procedimiento de Monte Carlo para las simulaciones. En estas simulaciones, los presentes inventores utilizaron la estimación del fabricante de la tasa de error de la polimerasa Phusion con un ajuste apropiado para la capacidad de Safe-SeqS para detectar mutaciones en una sola hebra (4). Obsérvese que los errores introducidos en el ciclo 19, así como en los dos ciclos de asignación de UID, darían lugar a cambios en solo una hebra del dúplex, es decir, darían lugar a un supermutante en lugar de a dos. En cada experimento, los presentes inventores asumieron que había una eficiencia constante de amplificación dada por el número total de moldes obtenidos al final del experimento (es decir, si el número de familias de UID era N, los presentes inventores supusieron que el número de moldes aumentó por un factor de $N/2^{19}$ en cada ciclo). Los presentes inventores ealizaron mil simulaciones para cada uno de los siete experimentos y los resultados se presentan en la tabla 4.

Referencias (solo para los ejemplos 8-11)

1. Herman DS, y otros (2009) Filter–based hybridization capture of subgenomes enables resequencing and copy–number detection. Nat Methods 6:507–510.
2. Vogelstein B & Kinzler KW (1999) Digital PCR. Proc Natl Acad Sci U S A 96:9236–9241.
3. Chory J & Pollard JD, Jr. (2001) Separation of small DNA fragments by conventional gel electrophoresis. Curr Protoc Mol Biol Chapter 2:Unit2 7.
4. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N–terminal deletion. Gene 112:29–35.
5. Collins AR (1999) Oxidative DNA damage, antioxidants, and cancer. Bioessays 21:238–246.
6. Morley AA, Cox S, & Holliday R (1982) Human lymphocytes resistant to 6–thioguanine increase with age. Mech Ageing Dev 19:21–26.
7. Trainor KJ, y otros (1984) Mutation frequency in human lymphocytes increases with age. Mech Ageing Dev 27:83–86.
8. Grist SA, McCarron M, Kutlaca A, Turner DR, & Morley AA (1992) In vivo human somatic mutation: frequency and spectrum with age. Mutat Res 266:189–196.
9. Williams GT, Geraghty JM, Campbell F, Appleton MA, & Williams ED (1995) Normal colonic mucosa in hereditary non–polyposis colorectal cancer shows no generalised increase in somatic mutation. Br J Cancer 71:1077–1080.
10. Campbell F, Appleton MA, Shields CJ, & Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right– and left–sided sporadic colorectal carcinomas. J Pathol 186:31–35.
11. Araten DJ, Nafa K, Pakdeesuwan K, & Luzzatto L (1999) Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. Proc Natl Acad Sci USA 96:5209–5214.
12. Araten DJ, y otros (2005) A quantitative measurement of the human somatic mutation rate. Cancer Res 65:8111–8117.
13. Monnat RJ, Jr. & Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. Proc Natl Acad Sci USA 82:2895–2899.
14. Bodenteich A, Mitchell LG, & Merril CR (1991) A lifetime of retinal light exposure does not appear to increase mitochondrial mutations. Gene 108:305–309.
15. Howell N, Kubacka I, & Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? Am J Hum Genet 59:501–509.
16. Khrapko K, y otros (1997) Mitochondrial mutational spectra in human cells and tissues. Proc Natl Acad Sci USA 94:13798–13803.
17. Heyer E, y otros (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep–rooting pedigrees. Am J Hum Genet 69:1113–1126.
18. Howell N, y otros (2003) The pedigree rate of sequence divergence in the human mitochondrial genome:

there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72:659-670.
19. Taylor RW, y otros (2003) Mitochondrial DNA mutations in human colonic crypt stem cells. J Clin Invest 112:1351–1360.
20. Luria SE & Delbruck M (1943) Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28:491–511.

Referencias (para todos excepto los ejemplos 8-11)

1. Luria SE & Delbruck M (1943) Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28:491–511.
2. Roach JC, y otros (2010) Analysis of genetic inheritance in a family quartet by whole–genome sequencing. Science 328:636-639.
3. Durbin RM, y otros (2010) A map of human genome variation from population–scale sequencing. Nature 467:1061–1073.
4. Shibata D (2011) Mutation and epigenetic molecular clocks in cancer. Carcinogenesis 32:123–128.
5. McMahon MA, y otros (2007) The HBV drug entecavir – effects on HIV–1 replication and resistance. N Engl J Med 356:2614–2621.
6. Eastman PS, y otros (1998) Maternal viral genotypic zidovudine resistance and infrequent failure of zidovudine therapy to prevent perinatal transmission of human immunodeficiency virus type 1 in pediatric AIDS Clinical Trials Group Protocol 076. J Infect Dis 177:557–564.
7. Chiu RW, y otros (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 105:20458–20463.
8. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, & Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 105:16266–16271.
9. Hoque MO, y otros (2003) High–throughput molecular analysis of urine sediment for the detection of bladder cancer by high–density single–nucleotide polymorphism array. Cancer Res 63:5723–5726.
10. Thunnissen FB (2003) Sputum examination for early detection of lung cancer. J Clin Pathol 56:805–810.
11. Diehl F, y otros (2008) Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. Gastroenterology 135:489–498.
12. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N–terminal deletion. Gene 112:29–35.
13. Araten DJ, y otros (2005) A quantitative measurement of the human somatic mutation rate. Cancer Res 65:8111–8117.
14. Campbell F, Appleton MA, Shields CJ, & Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right– and left–sided sporadic colorectal carcinomas. J Pathol 186:31–35.
15. Tindall KR & Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. Biochemistry 27:6008-6013.
16. Kunkel TA (1985) The mutational specificity of DNA polymerase–beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. J Biol Chem 260:5787–5796.
17. van Dongen JJ & Wolvers–Tettero IL (1991) Analysis of immunoglobulin and T cell receptor genes. Part II: Possibilities and limitations in the diagnosis and management of lymphoproliferative diseases and related disorders. Clin Chim Acta 198:93–174.
18. Grist SA, McCarron M, Kutlaca A, Turner DR, & Morley AA (1992) In vivo human somatic mutation: frequency and spectrum with age. Mutat Res 266:189–196.
19. Liu Q & Sommer SS (2004) Detection of extremely rare alleles by bidirectional pyrophosphorolysis–activated polymerization allele–specific amplification (Bi–PAP–A): measurement of mutation load in mammalian tissues. Biotechniques 36:156–166.
20. Monnat RJ, Jr. & Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. Proc Natl Acad Sci USA 82:2895–2899.
21. Shi C, y otros (2004) LigAmp for sensitive detection of single–nucleotide differences. Nat Methods 1:141–147.
22. Keohavong P & Thilly WG (1989) Fidelity of DNA polymerases in DNA amplification. Proc Natl Acad Sci USA 86:9253–9257.
23. Sidransky D, y otros (1991) Identification of p53 gene mutations in bladder cancers and urine samples. Science 252:706–709.
24. Bielas JH & Loeb LA (2005) Quantification of random genomic mutations. Nat Methods 2:285–290.
25. Vogelstein B & Kinzler KW (1999) Digital PCR. Proc Natl Acad Sci U S A 96:9236–9241.
26. Mitra RD, y otros (2003) Digital genotyping and haplotyping with polymerase colonies. Proc Natl Acad Sci USA 100:5926–5931.
27. Chetverina HV, Samatov TR, Ugarov VI, & Chetverin AB (2002) Molecular colony diagnostics: detection and quantitation of viral nucleic acids by in–gel PCR. Biotechniques 33:150–152, 154, 156.
28. Zimmermann BG, y otros (2008) Digital PCR: a powerful new tool for noninvasive prenatal diagnosis? Prenat Diagn 28:1087–1093.
29. Dressman D, Yan H, Traverso G, Kinzler KW, & Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 100:8817–8822.
30. Ottesen EA, Hong JW, Quake SR, & Leadbetter JR (2006) Microfluidic digital PCR enables multigene

analysis of individual environmental bacteria. Science 314:1464–1467.
31. Quail MA, y otros (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5:1005–1010.
32. Nazarian R, y otros (2010) Melanomas acquire resistance to B–RAF(V600E) inhibition by RTKorN–RAS upregulation. Nature 468:973–977.
33. He Y, y otros (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464:610-614.
34. Gore A, y otros (2011) Somatic coding mutations in human induced pluripotent stem cells. Nature 471:63-67.
35. Dohm JC, Lottaz C, Borodina T, & Himmelbauer H (2008) Substantial biases in ultrashort read data sets from high–throughput DNA sequencing. Nucleic Acids Res 36:e105.
36. Erlich Y, Mitra PP, delaBastide M, McCombie WR, & Hannon GJ (2008) Alta–Cyclic: a self–optimizing base caller for next–generation sequencing. Nat Methods 5:679-682.
37. Rougemont J, y otros (2008) Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 9:431.
38. Druley TE, y otros (2009) Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6:263–265.
39. Vallania FL, y otros (2010) High–throughput discovery of rare insertions and deletions in large cohorts. Genome Res 20:1711–1718.
40. McCloskey ML, Stoger R, Hansen RS, & Laird CD (2007) Encoding PCR products with batch–stamps and barcodes. Biochem Genet 45:761–767.
41. Parameswaran P, y otros (2007) A pyrosequencing–tailored nucleotide barcode design unveils opportunities for large–scale sample multiplexing. Nucleic Acids Res 35:e130.
42. Craig DW, y otros (2008) Identification of genetic variants using bar–coded multiplexed sequencing. Nat Methods 5:887–893.
43. Miner BE, Stoger RJ, Burden AF, Laird CD, & Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin–bisulfite PCR. Nucleic Acids Res 32:e135.
44. Herman DS, y otros (2009) Filter–based hybridization capture of subgenomes enables resequencing and copy–number detection. Nat Methods 6:507–510.
45. Jones PA & Baylin SB (2007) The epigenomics of cancer. Cell 128:683-692.
46. de Boer JG & Ripley LS (1988) An in vitro assay for frameshift mutations: hotspots for deletions of 1 bp by Klenow–fragment polymerase share a consensus DNA sequence. Genetics 118:181–191.
47. Eckert KA & Kunkel TA (1990) High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. Nucleic Acids Res 18:3739–3744.
48. Kosuri S, y otros (2010) Scalable gene synthesis by selective amplification of DNA pools from high–fidelity microchips. Nat Biotechnol 28:1295–1299.
49. Matzas M, y otros (2010) High–fidelity gene synthesis by retrieval of sequence–verified DNA identified using high–throughput pyrosequencing. Nat Biotechnol 28:1291–1294.
50. Li J, y otros (2008) Replacing PCR with COLD–PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. Nat Med 14:579–584.
51. Eid J, y otros (2009) Real–time DNA sequencing from single polymerase molecules. Science 323:133–138.

CLÁUSULAS

1. Un procedimiento para analizar secuencias de ácidos nucleicos, que comprende:

unir una secuencia de ácido nucleico de identificación única (UID) a un primer extremo de cada uno de una pluralidad de fragmentos de ácido nucleico de analito para formar fragmentos de ácido nucleico de analito identificados de forma única;
determinar de forma redundante la secuencia de nucleótidos de un fragmento de ácido nucleico de analito identificado de forma única, en la que determinadas secuencias de nucleótidos que comparten una UID forman una familia de miembros;
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ácido nucleico de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y la secuencia se encuentra en, como mínimo, dos familias.

2. El procedimiento de la cláusula 1, en el que antes de la etapa de determinar de forma redundante, se amplifican los fragmentos de ácido nucleico de analito identificados de forma única.

3. El procedimiento de la cláusula 1, en el que la secuencia de nucleótidos se identifica cuando el 50% de miembros de la familia contiene la secuencia.

4. El procedimiento de la cláusula 1, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 70% de miembros de la familia contiene la secuencia.

5. El procedimiento de la cláusula 1, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 90% de miembros de la familia contiene la secuencia.

6. El procedimiento de la cláusula 1, en el que la secuencia de nucleótidos se identifica cuando el 100% de los miembros de la familia contiene la secuencia.

7. El procedimiento de la cláusula 1, en el que la etapa de unión se realiza mediante reacción en cadena de la polimerasa.

8. El procedimiento de la cláusula 1, en el que un primer sitio de cebado universal está unido a un segundo extremo de cada uno de una pluralidad de fragmentos de ácido nucleico de analito.

9. El procedimiento de la cláusula 7, en el que, como mínimo, se realizan dos ciclos de reacción en cadena de la polimerasa, de manera que una familia se forma de fragmentos de ácido nucleico de analito identificados de forma única que tienen un UID en el primer extremo y un primer sitio de cebado universal en un segundo extremo.

10. El procedimiento de la cláusula 1, en el que el UID está unido covalentemente a un segundo sitio de cebado universal.

11. El procedimiento de la cláusula 8, en el que el UID está unido covalentemente a un segundo sitio de cebado universal.

12. El procedimiento de la cláusula 11, en el que antes de la etapa de determinar de forma redundante, los fragmentos de ácido nucleico de analito identificados de forma única se amplifican utilizando un par de cebadores que son complementarios del primer y segundo sitios de cebado universal, respectivamente.

13. El procedimiento de la cláusula 10, en el que el UID está unido al extremo 5' de un fragmento de ácido nucleico de analito y el segundo sitio de cebado universal está en 5' del UID.

14. El procedimiento de la cláusula 10, en el que el UID está unido al extremo 3' de un fragmento de ácido nucleico de analito y el segundo sitio de cebado universal está en 3' del UID.

15. El procedimiento de la cláusula 1, en el que los fragmentos de ácido nucleico de analito se forman aplicando una fuerza de cizallamiento al ácido nucleico de analito.

16. El procedimiento de la cláusula 7, en el que antes de la etapa de determinar de forma redundante, los fragmentos de ácido nucleico de analito identificados de forma única se someten a amplificación y en el que antes de dicha amplificación se utiliza una exonucleasa específica de una sola hebra para digerir los cebadores en exceso utilizados para unir el UID a los fragmentos de ácido nucleico de analito.

17. El procedimiento de la cláusula 16, en el que antes de la etapa de determinar de forma redundante, los fragmentos de ácido nucleico de analito identificados de manera única están sujetos a amplificación, y en el que antes de dicha amplificación, la exonucleasa específica de hebra única es inactivada, inhibida o eliminada.

18. El procedimiento de la cláusula 17, en el que la exonucleasa específica de hebra única se inactiva mediante tratamiento térmico.

19. El procedimiento de la cláusula 16, en el que los cebadores utilizados en dicha amplificación comprenden una o más modificaciones químicas que los hacen resistentes a las exonucleasas.

20. El procedimiento de la cláusula 16, en el que los cebadores utilizados en dicha amplificación comprenden uno o más enlaces fosforotioato.

21. El procedimiento para analizar secuencias de ácidos nucleicos, que comprende:

unir una secuencia identificadora única (UID) a un primer extremo de cada uno de una pluralidad de fragmentos de ADN de analito utilizando, como mínimo, dos ciclos de amplificación con un primer y segundo cebadores para formar fragmentos de ADN de analito identificados de forma única, en los que los UID están en exceso de los fragmentos de ADN de analito durante la amplificación, en el que los primeros cebadores comprenden:

- un primer segmento complementario a un amplicón deseado;
- un segundo segmento que contiene el UID;
- un tercer segmento que contiene un sitio de cebado universal para su posterior amplificación; y en el que los segundos cebadores comprenden un sitio de cebado universal para su posterior amplificación; en el que cada ciclo de amplificación une un sitio de cebado universal a una hebra;

amplificar los fragmentos de ADN de analito identificados de forma única para formar una familia de fragmentos de ADN de analito identificados de forma única a partir de cada fragmento de ADN de analito identificado de forma única; y

determinar la secuencia de nucleótidos de una pluralidad de miembros de una familia.

22. El procedimiento de la cláusula 21, en el que los segundos cebadores comprenden cada uno un UID.

23. El procedimiento de la cláusula 21, que comprende, además, las etapas de:

comparar secuencias de una familia de fragmentos de ADN de analito identificados de forma única; e
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y está presente en, como mínimo, dos familias.

24. El procedimiento de la cláusula 2, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 5% de miembros de la familia contiene la secuencia.

25. El procedimiento de la cláusula 23, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 25% de miembros de la familia contiene la secuencia.

26. El procedimiento de la cláusula 23, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 50% de miembros de la familia contiene la secuencia.

27. El procedimiento de la cláusula 2, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 70% de miembros de la familia contiene la secuencia.

28. El procedimiento de la cláusula 23, en el que la secuencia de nucleótidos se identifica cuando, como mínimo, el 90% de miembros de la familia contiene la secuencia.

29. El procedimiento de la cláusula 21, en el que los UID son de 2 a 4.000 bases, ambas cifras incluidas.

30. El procedimiento de la la cláusula 21, en el que antes de la etapa de amplificación de los fragmentos de ADN de analito identificados de forma única, se utiliza una exonucleasa específica de una sola hebra para digerir los cebadores en exceso utilizados para unir el UID a los fragmentos de ADN de analito.

31. Procedimiento de la cláusula 30, en el que antes de la etapa de amplificación, la exonucleasa específica de una sola hebra se inactiva, inhibe o elimina.

32. El procedimiento de la cláusula 31, en el que la exonucleasa específica de hebra única se inactiva mediante tratamiento térmico.

33. El procedimiento de la cláusula 30, en el que los cebadores utilizados en la etapa de amplificación comprenden uno o más enlaces fosforotioato.

34. Un procedimiento para analizar ADN utilizando secuencias identificadoras únicas (UID) endógenas, que comprende:

unir oligonucleótidos adaptadores a extremos de fragmentos de ADN de analito de entre 30 y 2.000 bases, ambas incluidas, para formar fragmentos adaptados, en los que cada extremo de un fragmento antes de dicha unión es un UID endógeno para el fragmento;
amplificar los fragmentos adaptados utilizando cebadores complementarios a los oligonucleótidos adaptadores para formar familias de fragmentos adaptados;
determinar la secuencia de nucleótidos de una pluralidad de miembros de una familia; comparar secuencias de nucleótidos de la pluralidad de miembros de la familia; e
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 50 4% de los miembros de la familia contiene la secuencia y la secuencia se encuentra en, como mínimo, dos familias.

35. El procedimiento de la cláusula 34, que comprende además:

enriquecer para fragmentos que representan uno o más genes seleccionados mediante la captura de un subconjunto de los fragmentos utilizando oligonucleótidos de captura complementarios a genes seleccionados en el ADN de analito.

36. El procedimiento de la cláusula 34, que comprende además:

enriquecer para fragmentos que representan uno o más genes seleccionados mediante amplificación de fragmentos complementarios a genes seleccionados.

37. El procedimiento de la cláusula 35 ó 36, en el que la etapa de unión es anterior a la etapa de enriquecimiento.

38. Procedimiento de la cláusula 34, en el que los fragmentos se forman mediante cizallamiento.

39. Una población de pares de cebadores, en la que cada par comprende un primer y segundo cebadores para amplificar e identificar un gen o parte de gen, en la que:

- el primer cebador comprende una primera parte de 10-100 nucleótidos complementarios al gen o parte del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un tercer cebador;
- el segundo cebador comprende una primera parte de 10-100 nucleótidos complementarios a la parte del gen o del gen y una segunda parte de 10 a 100 nucleótidos que comprende un sitio para la hibridación con un cuarto cebador, en el que interpuesta entre la primera parte y la segunda parte del segundo cebador hay una tercera parte que consiste en de 2 a 4.000 nucleótidos que forman un identificador único (UID):
- en el que los identificadores únicos en la población tienen, como mínimo, 4 secuencias diferentes, en el que el primer y segundo cebadores son complementarios u opuestos a las hebras del gen o parte génica.

40. El procedimiento de la cláusula 39, en el que el primer cebador comprende, además, un identificador único (UID).

41. La población de la cláusula 39, en la que los identificadores únicos en la población tienen, como mínimo, como mínimo, 16, como mínimo, 64, como mínimo, 256, como mínimo, 1.024, como mínimo, 4.096, como mínimo, 16.384, como mínimo, 65.536, como mínimo, 262.144, como mínimo, 1.048.576, como mínimo, 4.194.304, como mínimo, 16.777.216, o, como mínimo, 67.108.864 secuencias diferentes.

42. Un kit que comprende la población de cebadores de la cláusula 39 y el tercer y cuarto cebadores complementarios a las segundas partes de cada uno del primer y segundo cebadores.

43. La población de la cláusula 39, en la que el UID comprende secuencias seleccionadas aleatoriamente.

44. La población de la cláusula 39, en la que el UID comprende secuencias de nucleótidos predefinidos.

45. La población de la cláusula 39, en la que el UID comprende tanto secuencias seleccionadas aleatoriamente como nucleótidos predefinidos.

46. El procedimiento de la cláusula 2, 21 ó 34 6, en el que antes de la amplificación, el ADN de analito se trata con bisulfito para convertir bases de citosina no metiladas en uracilo.

47. El procedimiento de la cláusula 1, 21 ó 34 que comprende, además, la etapa de comparar el número de familias que representan un primer fragmento de ADN de analito con el número de familias que representan un segundo fragmento de ADN de analito para determinar una concentración relativa de un primer fragmento de ADN de analito a un segundo fragmento de ADN de analito en la pluralidad de fragmentos de ADN de analito.

LISTADO DE SECUENCIAS

<110> Vogelstein, Bert
Kinzler, Kenneth W.
Papadopoulos, Nickolas
Kinde, Isaac

<120> Sistema de secuenciación segura

<130> 001107,00873

<160> 81

<170> FastSEQ for Windows Versión 4.0

<210> 1
<211> 32
<212> ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 1

gatcggaaga gcggttcagc aggaatgccg ag 32

<210> 2
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 2
acactctttc cctacacgac gctcttccga tct 33

<210> 3
<211> 62
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 3

aatgatacgg cgaccaccga gatctacaca cactctttcc ctacacgacg ctcttccgat 60
ct 62

<210> 4
<211> 57
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 4
caagcagaag acggcatacg agatctcggc attcctgctg aaccgctctt ccgatct 57

<210> 5
<211> 62
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 5

aatgatacgg cgaccaccga gatctacaca cactctttcc ctacacgacg ctcttccgat 60
ct 62

<210> 6
<211> 57
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 6
caagcagaag acggcatacg agatctcggc attcctgctg aaccgctctt ccgatct 57

<210> 7
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 7
acactctttc cctacacgac gctcttccga tct 33

<210> 8
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 8
ctcggcattc ctgctgaacc gctcttccga tct 33

<210> 9
<211> 32
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 9
gatcggaaga gcggttcagc aggaatgccg ag 32

<210> 10
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 10
acactctttc cctacacgac gctcttccga tct 33

<210> 11
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 11
cgtgatacac tctttcccta cacgacgctc ttccgatct 39

<210> 12
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 12
acatcgacac tctttcccta cacgacgctc ttccgatct 39

<210> 13
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 13
gcctaaacac tctttcccta cacgacgctc ttccgatct 39

<210> 14
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 14
tggtcaacac tctttcccta cacgacgctc ttccgatct 39

<210> 15
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 15
cactgtacac tctttcccta cacgacgctc ttccgatct 39

<210> 16
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 16
attggcacac tctttcccta cacgacgctc ttccgatct 39

<210> 17
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 17
gatctgacac tctttcccta cacgacgctc ttccgatct 39

<210> 18
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 18
tcaagtacac tctttcccta cacgacgctc ttccgatct 39

<210> 19
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 19
ctgatcacac tctttcccta cacgacgctc ttccgatct 39

<210> 20
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 20
aagctaacac tctttcccta cacgacgctc ttccgatct 39

<210> 21
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 21
gtagccacac tctttcccta cacgacgctc ttccgatct 39

<210> 22
<211> 39
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 22
tacaagacac tctttcccta cacgacgctc ttccgatct 39

<210> 23
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 23
ctcggcattc ctgctgaacc gctcttccga tct 33

<210> 24
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 24
aatgatacgg cgaccaccga gatctacacc agcaggcctt ataataaaaa taatga 56

<210> 25
<211> 51
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 25
caagcagaag acggcatacg agattgactg aatataaact tgtggtagtt g 51

<210> 26
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 26
acactctttc cctacacgac gctcttccga tct 33

<210> 27
<211> 33
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 27
ctcggcattc ctgctgaacc gctcttccga tct 33

<210> 28
<211> 29
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 28
cggaagagcg tcgtgtaggg aaagagtgt 29

<210> 29
<211> 29
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 29
cggaagagcg gttcagcagg aatgccgag 29

<210> 30
<211> 23
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 30
ggttacaggc tcatgatgta acc 23

<210> 31
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 31
gataccagct tggtaatggc a 21

<210> 32
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores
<221> misc_feature
<222> (1)...(56)
<223> n = A, T, C o G

<400> 32
cgacgtaaaa cgacggccag tnnnnnnnnn nnnggttaca ggctcatgat gtaacc 56

<210> 33
<211> 46
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 33
cacacaggaa acagctatga ccatggatac cagcttggta atggca 46

<210> 34
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 34
aatgatacgg cgaccaccga gatctacacc gtgatcgacg taaaacgacg gccagt 56

<210> 35
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 35
aatgatacgg cgaccaccga gatctacaca catcgcgacg taaaacgacg gccagt 56

<210> 36
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 36
aatgatacgg cgaccaccga gatctacacg cctaacgacg taaaacgacg gccagt 56

<210> 37
<211> 56

<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 37
aatgatacgg cgaccaccga gatctacact ggtcacgacg taaaacgacg gccagt 56

<210> 38
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 38
aatgatacgg cgaccaccga gatctacacc actgtcgacg taaaacgacg gccagt 56

<210> 39
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 39
aatgatacgg cgaccaccga gatctacaca ttggccgacg taaaacgacg gccagt 56

<210> 40
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 40
aatgatacgg cgaccaccga gatctacacg atctgcgacg taaaacgacg gccagt 56

<210> 41
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 41
aatgatacgg cgaccaccga gatctacact caagtcgacg taaaacgacg gccagt 56

<210> 42
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 42
aatgatacgg cgaccaccga gatctacacc tgatccgacg taaaacgacg gccagt 56

<210> 43
<211> 56

<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 43
aatgatacgg cgaccaccga gatctacaca agctacgacg taaaacgacg gccagt 56

<210> 44
<211> 49
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 44
caagcagaag acggcatacg agatcacaca ggaaacagct atgaccatg 49

<210> 45
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 45
cgacgtaaaa cgacggccag t 21

<210> 46
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 46
actggccgtc gttttacgtc g 21

<210> 47
<211> 58
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores
<221> misc_feature
<222> (1)...(58)
<223> n = A, T, C o G

<400> 47
cgacgtaaaa cgacggccag tnnnnnnnnn nnnnngcagc aacagtctta cctggact 58

<210> 48
<211> 48
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 48
cacacaggaa acagctatga ccatgtccac atcctcttcc tcaggatt 48

<210> 49
<211> 50
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 49
aatgatacgg cgaccaccga gatctacacc gacgtaaaac gacggccagt 50

<210> 50
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 50
caagcagaag acggcatacg agatatcacg cacacaggaa acagctatga ccatg 55

<210> 51
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 51
caagcagaag acggcatacg agatcgatgt cacacaggaa acagctatga ccatg 55

<210> 52
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 52
caagcagaag acggcatacg agattgacca cacacaggaa acagctatga ccatg 55

<210> 53
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 53
caagcagaag acggcatacg agatgccaat cacacaggaa acagctatga ccatg 55

<210> 54
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 54
caagcagaag acggcatacg agatcagatc cacacaggaa acagctatga ccatg 55

<210> 55
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 55
caagcagaag acggcatacg agatacttga cacacaggaa acagctatga ccatg 55

<210> 56
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 56
caagcagaag acggcatacg agatgatcag cacacaggaa acagctatga ccatg 55

<210> 57
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 57
caagcagaag acggcatacg agattagctt cacacaggaa acagctatga ccatg 55

<210> 58
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 58
caagcagaag acggcatacg agatggctac cacacaggaa acagctatga ccatg 55

<210> 59
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 59
caagcagaag acggcatacg agatcttgta cacacaggaa acagctatga ccatg 55

<210> 60
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 60
cgacgtaaaa cgacggccag t 21

<210> 61
<211> 25
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 61
catggtcata gctgtttcct gtgtg 25

<210> 62
<211> 57
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores
<221> misc_feature
<222> (1)...(57)
<223> n = A, T, C o G

<400> 62
cgacgtaaaa cgacggccag tnnnnnnnnn nnnnnttacc gagaaagctc acaagaa 57

<210> 63
<211> 45
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 63
cacacaggaa acagctatga ccatgatgct aaggcgagga tgaaa 45

<210> 64
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 64
aatgatacgg cgaccaccga gatctacaca catcgcgacg taaaacgacg gccagt 56

<210> 65
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 65
aatgatacgg egaccaccga gatctacacg cctaacgacg taaaacgacg gccagt 56

<210> 66
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 66
aatgatacgg cgaccaccga gatctacact ggtcacgacg taaaacgacg gccagt 56

<210> 67
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 67
aatgatacgg cgaccaccga gatctacaca ttggccgacg taaaacgacg gccagt 56

<210> 68
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 68
aatgatacgg cgaccaccga gatctacacg atctgcgacg taaaacgacg gccagt 56

<210> 69
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 69
aatgatacgg cgaccaccga gatctacact caagtcgacg taaaacgacg gccagt 56

<210> 70
<211> 56
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 70
aatgatacgg cgaccaccga gatctacacc tgatccgacg taaaacgacg gccagt 56

<210> 71
<211> 49
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 71
caagcagaag acggcatacg agatcacaca ggaaacagct atgaccatg 49

<210> 72
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 72
cgacgtaaaa cgacggccag t 21

<210> 73
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 73
cctaattccc cccatcctta c 21

<210> 74
<211> 21
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 74
actggccgtc gttttacgtc g 21

<210> 75
<211> 77
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 75

ggttacaggc tcatgatgta acctctgtgt cttggtgtaa ctttaaaaca tatttttgcc 60
attaccaagc tggtatc 77

<210> 76
<211> 77
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 76

ggttacaggc tcatgatgta acctctgtgt cttggtgsaa ctttaaaaca tatttttgcc 60
attaccaagc tggtatc 77

<210> 77
<211> 55
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores
<221> misc_feature
<222> (1)...(55)
<223> n = A, T, C o G

<400> 77
acactctttc cctacacgac gctcnnnnnn nnnnnnggtg agtctgtgca ggcat 55

<210> 78
<211> 45
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 78
ctcgagcact gtcctgactg agacgatacc agcttggtaa tggca 45

<210> 79
<211> 59
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 79
aatgatacgg cgaccaccga gatctacacc gtgatacact ctttccctac acgacgctc 59

<210> 80
<211> 48
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 80
caagcagaag acggcatacg agatctcgag cactgtcctg actgagac 48

<210> 81
<211> 24
<212>ADN
<213> Secuencia artificial

<220>
<223> Cebadores y adaptadores

<400> 81
acactctttc cctacacgac gctc 24

REIVINDICACIONES

1. Procedimiento para identificar mutaciones por sustitución, inserción y deleción de una sola base en un fragmento de ácido nucleico de analito, que comprende:

unir una secuencia de ácido nucleico de identificación única (UID) a un primer extremo de cada uno de una pluralidad de fragmentos de ácido nucleico de analito para formar fragmentos de ácido nucleico de analito identificados de forma única;
determinar de forma redundante la secuencia de nucleótidos de un fragmento de ácido nucleico de analito identificado de forma única, en la que determinadas secuencias de nucleótidos que comparten una UID forman una familia de miembros;
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ácido nucleico de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y la secuencia se encuentra en, como mínimo, dos familias; e
identificar una mutación de sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito cuando la secuencia de nucleótidos que representa con precisión el fragmento de ácido nucleico de analito es diferente de una secuencia esperada en una sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito.

2. Procedimiento, según la reivindicación 1, en el que:

(1) antes de la etapa de determinación de forma redundante, se amplifican los fragmentos de ácido nucleico de analito identificados de forma única y en el que, opcionalmente, antes de la amplificación, el ADN de analito se trata con bisulfito para convertir bases de citosina no metilada en uracilo;
(2) la secuencia de nucleótidos se identifica cuando, como mínimo, el 70% de miembros de la familia contiene la secuencia;
(3) la secuencia de nucleótidos se identifica cuando, como mínimo, el 90% de miembros de la familia contiene la secuencia;
(4) la secuencia de nucleótidos se identifica cuando el 100% de miembros de la familia contiene la secuencia; o
(5) los fragmentos de ácido nucleico de analito se forman aplicando una fuerza de cizallamiento al ácido nucleico de analito.

3. Procedimiento, según la reivindicación 1, en el que la etapa de unión se realiza mediante reacción en cadena de la polimerasa.

4. Procedimiento, según la reivindicación 1, en el que un primer sitio de cebado universal está unido a un segundo extremo de cada uno de una pluralidad de fragmentos de ácido nucleico de analito, en el que, opcionalmente, el UID está unido covalentemente a un segundo sitio de cebado universal y, preferentemente, en el que antes de la etapa de determinación de forma redundante, los fragmentos de ácido nucleico de analito identificados de forma única se amplifican utilizando un par de cebadores que son complementarios del primer y segundo sitios de cebado universal, respectivamente.

5. Procedimiento, según la reivindicación 3, en el que, como mimo, se realizan dos ciclos de reacción en cadena de la polimerasa, de manera que una familia se forma de fragmentos de ácido nucleico de analito identificados de forma única que tienen un UID en el primer extremo y un primer sitio de cebado universal en un segundo extremo.

6. procedimiento, según la reivindicación 1, en el que el UID está unido covalentemente a un segundo sitio de cebado universal, en el que, opcionalmente:

(1) el UID está unido al extremo 5' de un fragmento de ácido nucleico de analito y el segundo sitio de cebado universal está en 5' del UID; o
(2) el UID está unido al extremo 3' de un fragmento de ácido nucleico de analito y el segundo sitio de cebado universal está en 3' del UID.

7. Procedimiento, según la reivindicación 3, en el que antes de la etapa de determinar de forma redundante, los fragmentos de ácido nucleico de analito identificados de forma única se someten a amplificación y en el que antes de dicha amplificación se utiliza una exonucleasa específica de una sola hebra para digerir los cebadores en exceso utilizados para unir el UID a los fragmentos de ácido nucleico de analito, en el que, opcionalmente:

(1) antes de la etapa de determinación de forma redundante, los fragmentos de ácido nucleico de analito identificados de manera única están sujetos a amplificación, y en el que antes de dicha amplificación, la exonucleasa específica de hebra única se inactiva, inhibe o elimina, en el que, preferentemente, la exonucleasa específica de hebra única se inactiva mediante tratamiento térmico;
(2) los cebadores utilizados en dicha amplificación comprenden una o más modificaciones químicas que los convierten en resistentes a las exonucleasas; o

(3) los cebadores utilizados en dicha amplificación comprenden uno o más enlaces fosforotioato.

8. Procedimiento para identificar mutaciones por sustitución, inserción y deleción de una sola base en un fragmento de ácido nucleico de analito, que comprende:

unir una secuencia identificadora única (UID) a un primer extremo de cada uno de una pluralidad de fragmentos de ADN de analito utilizando, como mínimo, dos ciclos de amplificación con primer y segundo cebadores para formar fragmentos de ADN de analito identificados de forma única, en los que los UID están en exceso de los fragmentos de ADN de analito durante la amplificación, en el que los primeros cebadores comprenden:

- un primer segmento complementario a un amplicón deseado;
- un segundo segmento que contiene el UID;
- un tercer segmento que contiene un sitio de cebado universal para la posterior amplificación;

y en el que los segundos cebadores comprenden un sitio de cebado universal para su posterior amplificación; en el que cada ciclo de amplificación une un sitio de cebado universal a una hebra;
amplificar los fragmentos de ADN de analito identificados de forma única para formar una familia de fragmentos de ADN de analito identificados de forma única a partir de cada fragmento de ADN de analito identificado de forma única; y
determinar secuencias de nucleótidos de una pluralidad de miembros de la familia, comprendiendo, además, el procedimiento las etapas de:

comparar secuencias de una familia de fragmentos de ADN de analito identificados de forma única;
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y la secuencia se encuentra en, como mínimo, dos familias; e
identificar una mutación de sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito cuando la secuencia de nucleótidos que representa con precisión el fragmento de ácido nucleico de analito es diferente de la secuencia esperada en una sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito, en el que, opcionalmente:

(1) la secuencia de nucleótidos se identifica cuando, como mínimo, el 70% de miembros de la familia contiene la secuencia; o
(2) la secuencia de nucleótidos se identifica cuando, como mínimo, el 90% de miembros de la familia contiene la secuencia.

9. Procedimiento, según la reivindicación 8, en el que:

(1) los segundos cebadores comprenden cada uno un UID; o
(2) Los UID tienen de 2 a 4.000 bases, ambas cifras incluidas.

10. Procedimiento, según la reivindicación 8, en el que antes de la etapa de amplificación de los fragmentos de ADN de analito identificados de forma única, se utiliza una exonucleasa específica de una sola hebra para digerir los cebadores en exceso utilizados para unir el UID a los fragmentos de ADN de analito, en el que, opcionalmente:

(1) antes de la etapa de amplificación, la exonucleasa específica de hebra única se inactiva, inhibe o elimina, en el que, preferentemente, la exonucleasa específica de hebra única se inactiva mediante tratamiento térmico; o
(2) los cebadores utilizados en la etapa de amplificación comprenden uno o más enlaces fosforotioato.

11. Procedimiento para identificar mutaciones por sustitución, inserción y deleción de una sola base en un fragmento de ácido nucleico de analito utilizando secuencias de identificación única (UID) endógenas, que comprende:

unir oligonucleótidos adaptadores a extremos de fragmentos de ADN de analito de entre 30 y 2.000 bases, ambas incluidas, para formar fragmentos adaptados, en los que cada extremo de un fragmento antes de dicha unión es un UID endógeno para el fragmento;
amplificar los fragmentos adaptados utilizando cebadores complementarios a los oligonucleótidos adaptadores para formar familias de fragmentos adaptados;
determinar la secuencia de nucleótidos de una pluralidad de miembros de una familia.
comparar las secuencia de nucleótidos de la pluralidad de miembros de una familia;
identificar una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 50% de los miembros de la familia contiene la secuencia y la secuencia se encuentra en, como mínimo, dos familias; e
identificar una mutación de sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito cuando la secuencia de nucleótidos que representa con precisión el fragmento de ácido nucleico de analito es diferente de la secuencia esperada en una sustitución, inserción o deleción de una sola base en el fragmento de ácido nucleico de analito.

12. Procedimiento, según la reivindicación 11, en el que:

(1) el procedimiento comprende, además, el enriquecimiento de fragmentos que representan uno o más genes seleccionados mediante la captura de un subconjunto de los fragmentos utilizando oligonucleótidos de captura complementarios a genes seleccionados en el ADN de analito y en el que, opcionalmente, la etapa de unión es anterior a la etapa de enriquecimiento;
(2) el procedimiento comprende, además, el enriquecimiento de fragmentos que representan uno o más genes seleccionados por medio de amplificación de fragmentos complementarios a genes seleccionados y, en el que, opcionalmente, la etapa de unión es anterior a la etapa de enriquecimiento;
(3) los fragmentos se forman mediante cizallamiento; o
(4) se identifica una secuencia de nucleótidos que representa con precisión un fragmento de ADN de analito cuando, como mínimo, el 70% de los miembros de la familia contiene la secuencia.

13. Procedimiento, según la reivindicación 8 u 11, en el que antes de la amplificación, el ADN de analito se trata con bisulfito para convertir bases de citosina no metilada en uracilo.

14. Procedimiento, según la reivindicación 1, 8 u 11 que comprende, además, la etapa de comparar el número de familias que representan un primer fragmento de ADN de analito con el número de familias que representan un segundo fragmento de ADN de analito para determinar una concentración relativa de un primer fragmento de ADN de analito a un segundo analito fragmento de ADN en la pluralidad de fragmentos de ADN de analito.

15. Procedimiento, según la reivindicación 1, 8 u 11, en el que se identifica que una secuencia de nucleótidos representa con precisión un fragmento de ADN de analito cuando

(a) la secuencia está presente en solo una de dos hebras del fragmento de ADN de analito; o

(b) la secuencia está presente en las dos hebras del fragmento de ADN de analito.

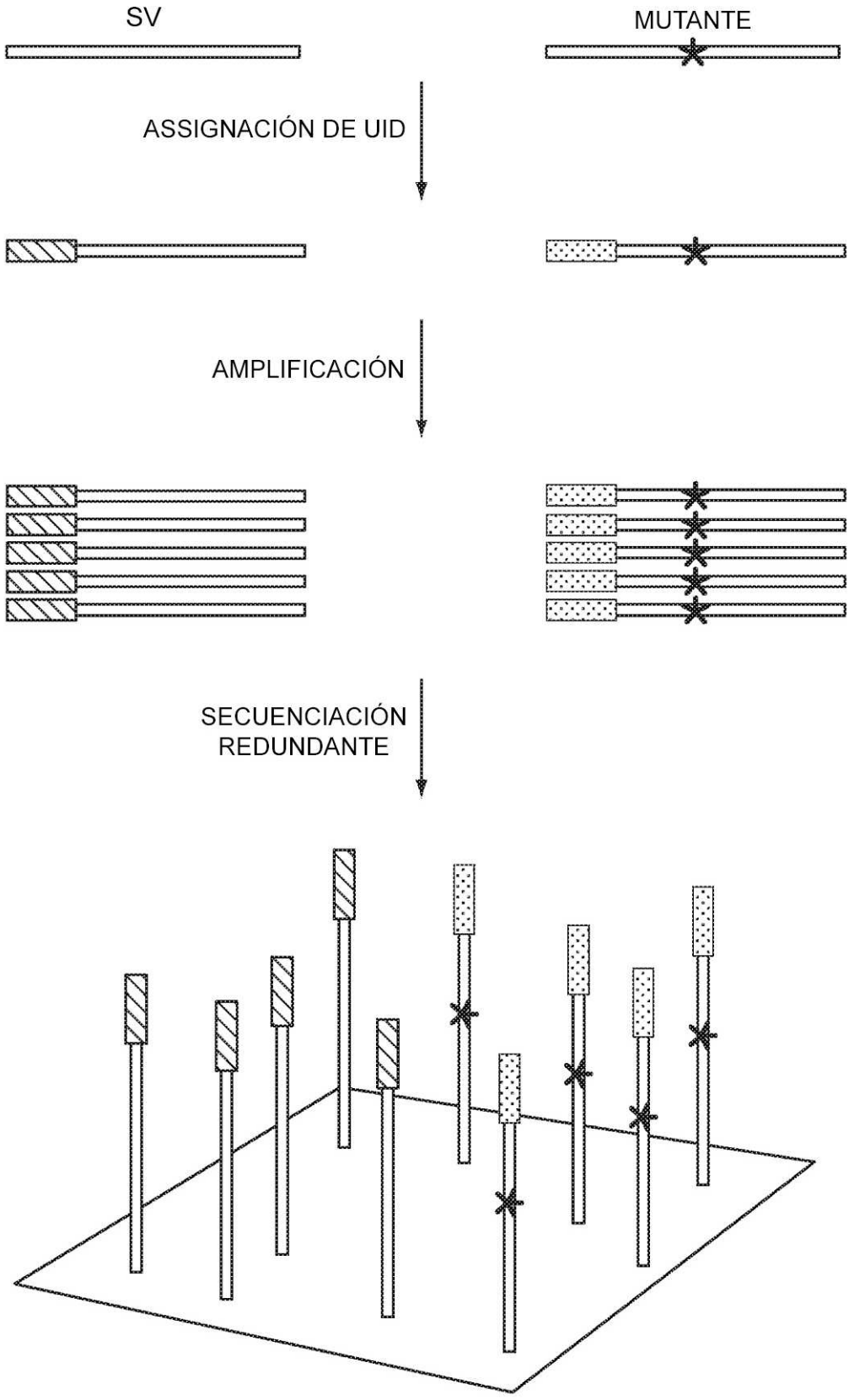
SV
MUTANTE
ASSIGNACIÓN DE UID
AMPLIFICACIÓN
SECUENCIACIÓN
REDUNDANTE

FIG. 1

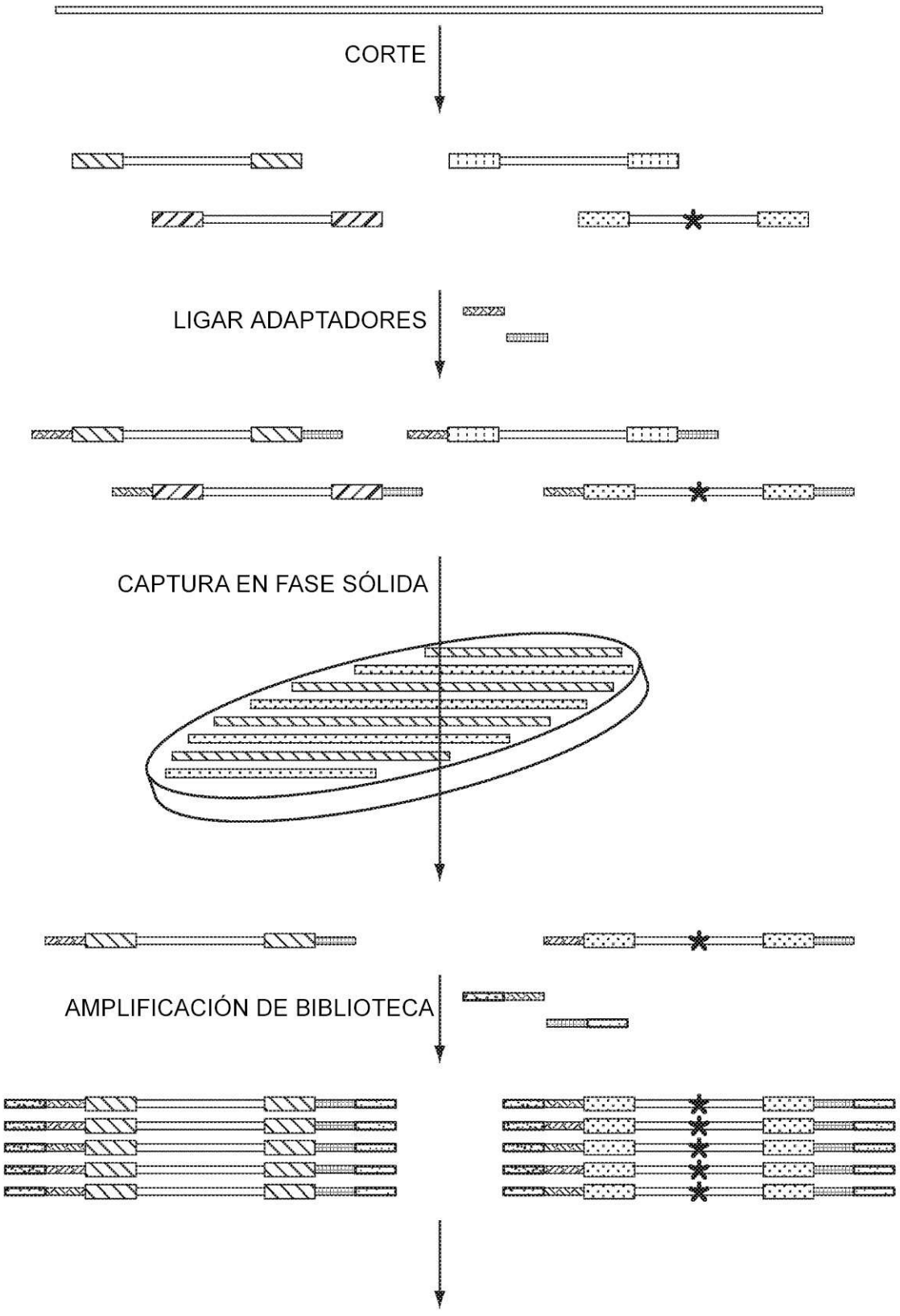
CORTE
LIGAR ADAPTADORES
CAPTURA EN FASE SÓLIDA
AMPLIFICACIÓN DE BIBLIOTECA
SECUENCIACIÓN REDUNDANTE

FIG. 2

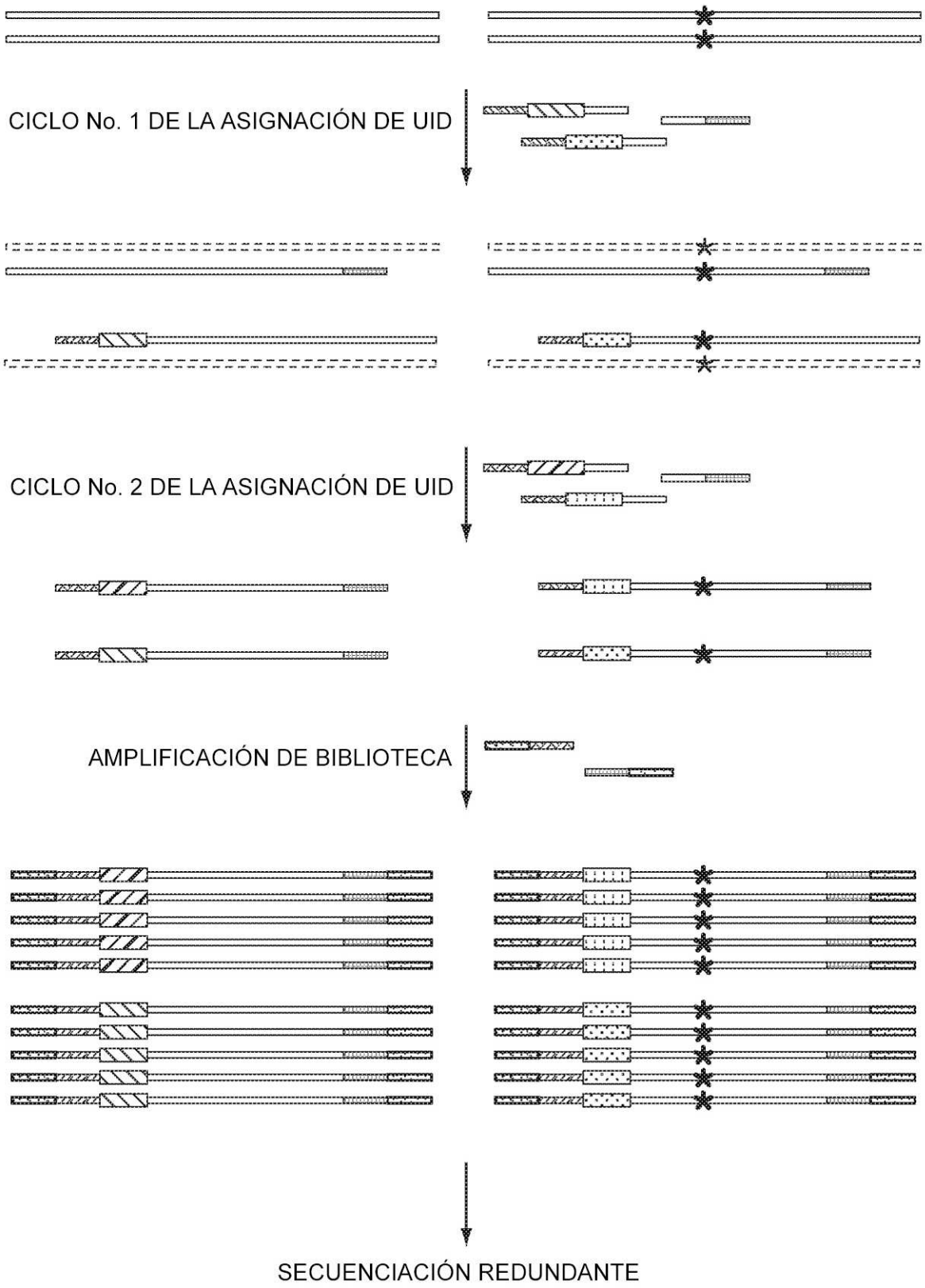
CICLO No. 1 DE LA ASIGNACIÓN DE UID
CICLO No. 2 DE LA ASIGNACIÓN DE UID
AMPLIFICACIÓN DE BIBLIOTECA
SECUENCIACIÓN REDUNDANTE

FIG. 3

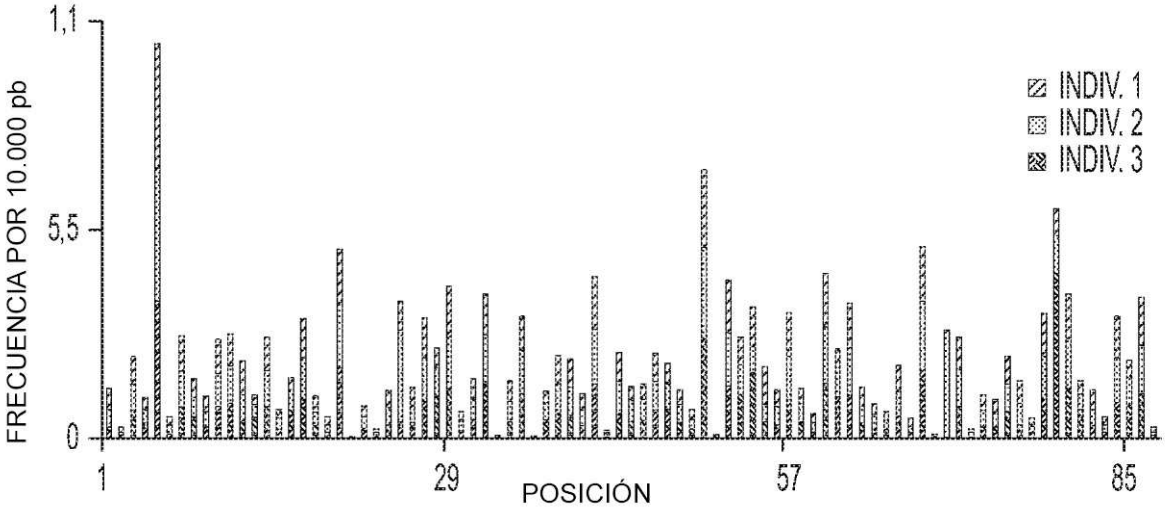
1,1
5,5
0
FRECUENCIA POR 10.000 pb
INDIV. 1
INDIV. 2
INDIV. 3
1
29
57
85
POSICIÓN

FIG. 4A

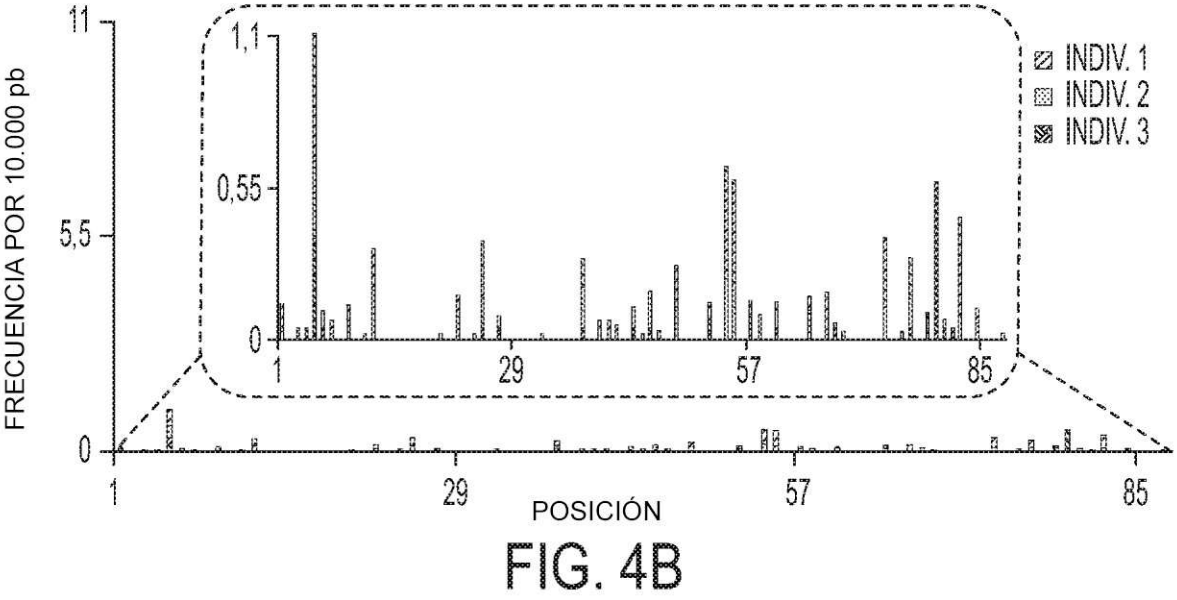
11
5,5
0
FRECUENCIA POR 10.000 pb
1,1
0,55
0
INDIV. 1
INDIV. 2
INDIV. 3
1
29
57
85
POSICIÓN

FIG. 4B

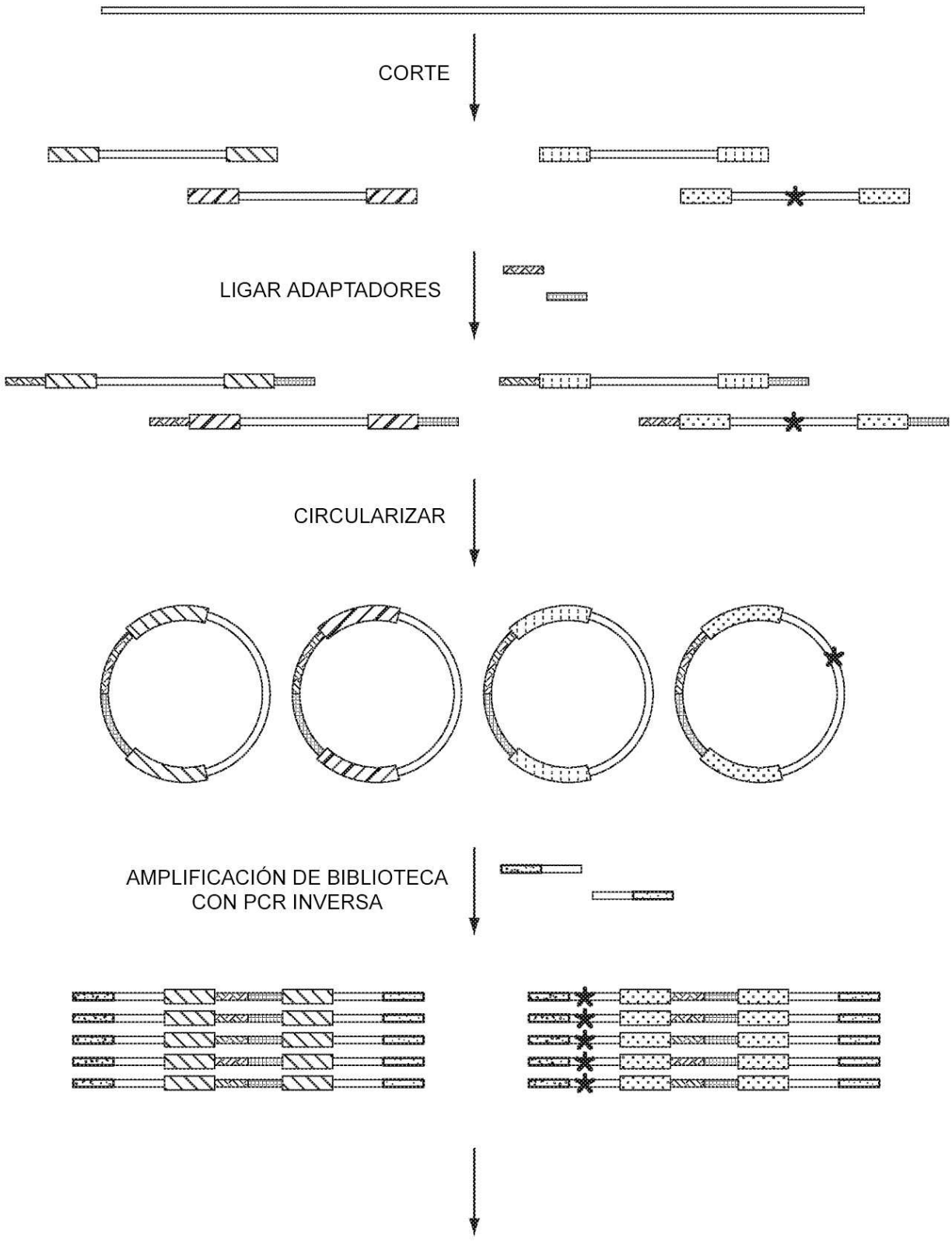
CORTE
LIGAR ADAPTADORES
CIRCULARIZAR
AMPLIFICACIÓN DE BIBLIOTECA
CON PCR INVERSA
SECUENCIACIÓN REDUNDANTE

FIG. 5

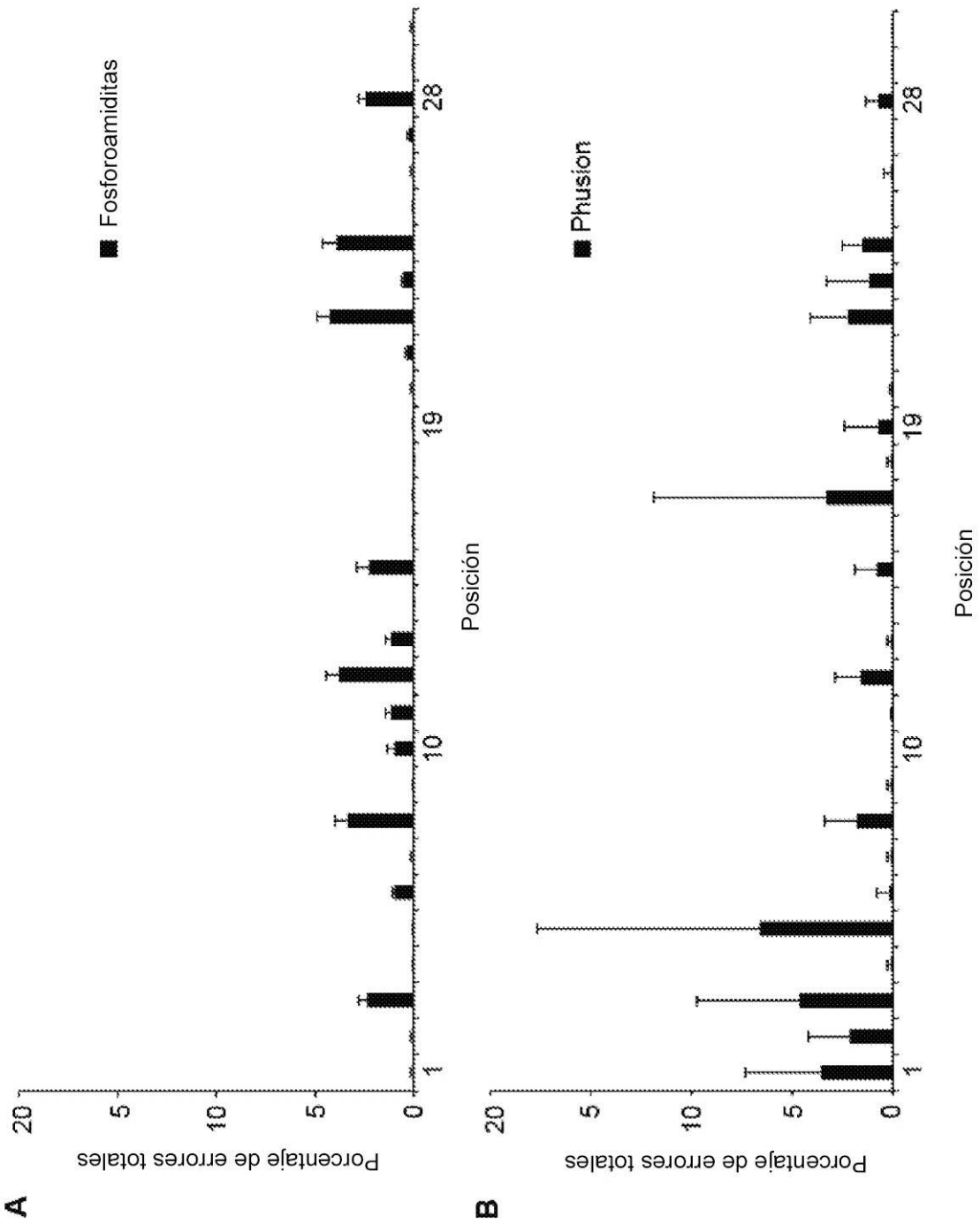
A
Fosforoamiditas
Porcentaje de errores totales
20
5
10
5
0
1
10
19
28
Posición
B
Phusion
Porcentaje de errores totales
20
5
10
5
0
1
10
19
28
Posición

Fig 6A-6B.

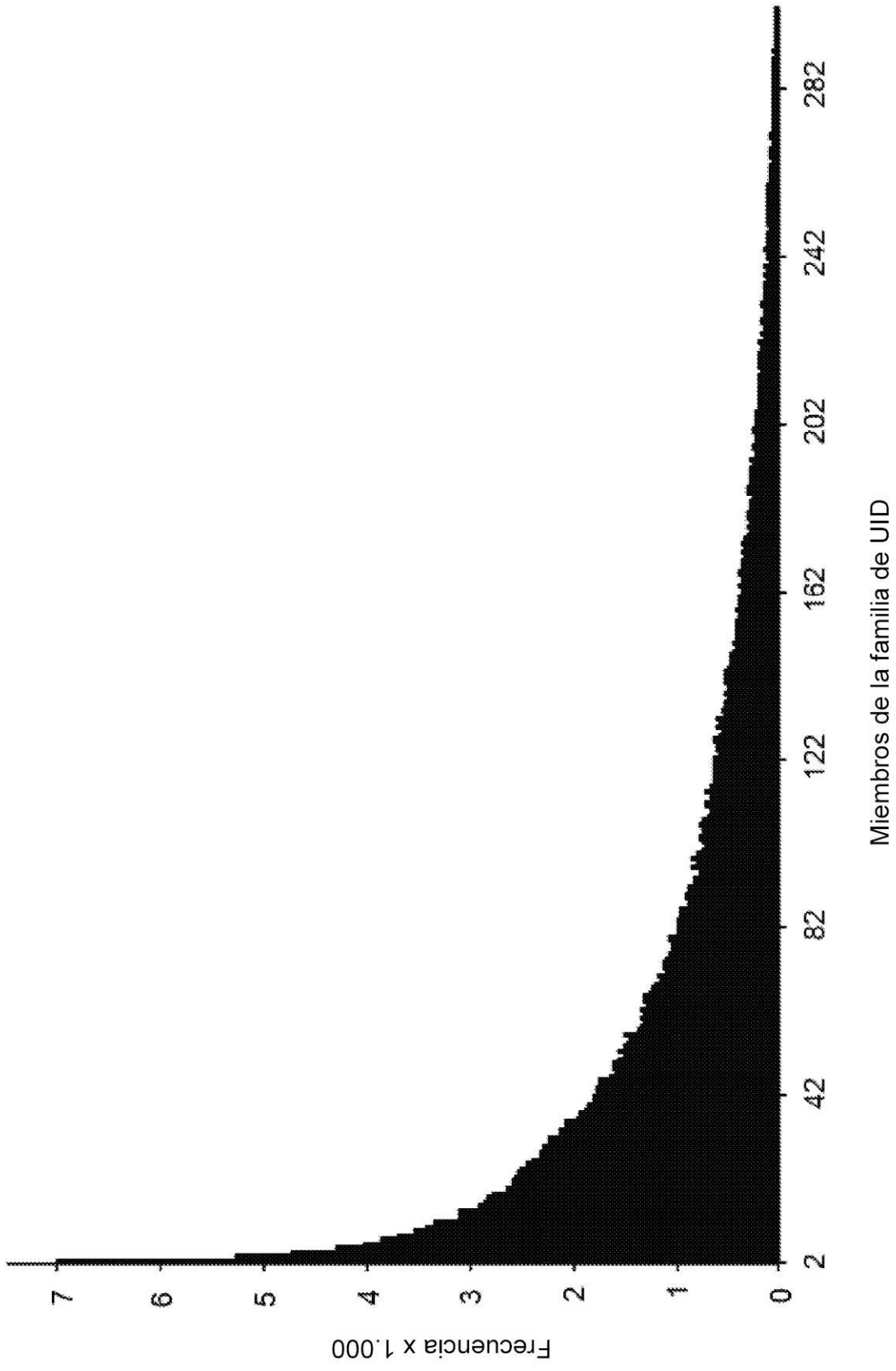
Frecuencia x 1.000
0
1
2
3
4
5
6
7
2
42
82
122
162
202
242
282
Miembros de la familia de UID

Fig. 7.