(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 701 749**

(51) Int. CI.:

C12N 15/10 (2006.01)
C12N 9/22 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **12.12.2013 PCT/US2013/074812**

(87) Fecha y número de publicación internacional: **19.06.2014 WO14093709**

(96) Fecha de presentación y número de la solicitud europea: **12.12.2013 E 13812447 (4)**

(97) Fecha y número de publicación de la concesión europea: **12.09.2018 EP 2931892**

(54) Título: **Métodos, modelos, sistemas y aparatos para identificar secuencias diana para enzimas Cas o sistemas CRISPR-Cas para secuencias diana y transmitir resultados de los mismos**

(30) Prioridad:

12.12.2012 US 201261736527 P
02.01.2013 US 201361748427 P
30.01.2013 US 201361758468 P
25.02.2013 US 201361769046 P
15.03.2013 US 201361791409 P
15.03.2013 US 201361802174 P
28.03.2013 US 201361806375 P
20.04.2013 US 201361814263 P
06.05.2013 US 201361819803 P
28.05.2013 US 201361828130 P
17.06.2013 US 201361835931 P
17.06.2013 US 201361836080 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
25.02.2019

(73) Titular/es:

THE BROAD INSTITUTE, INC. (33.3%)
415 Main Street
Cambridge, MA 02142, US;
MASSACHUSETTS INSTITUTE OF TECHNOLOGY (33.3%) y
PRESIDENT AND FELLOWS OF HARVARD COLLEGE (33.3%)

(72) Inventor/es:

ZHANG, FENG;
LI, YINQING;
SCOTT, DAVID, ARTHUR;
WEINSTEIN, JOSHUA, ASHER y
HSU, PATRICK

(74) Agente/Representante:

VALLEJO LÓPEZ, Juan Pedro

DESCRIPCIÓN

Métodos, modelos, sistemas y aparatos para identificar secuencias diana para enzimas Cas o sistemas CRISPR-Cas para secuencias diana y transmitir resultados de los mismos.

Campo de la invención

La presente divulgación se refiere en general a la ingeniería y optimización de los sistemas, métodos y composiciones utilizados para el control de la expresión génica que involucran el direccionamiento de secuencias, como la perturbación del genoma o la edición de genes, que se relacionan con las Repeticiones Palindrómicas Cortas Agrupadas y Regularmente Interespaciadas (CRISPR) y sus componentes.

Antecedentes de la invención

El sistema CRISPR/Cas o el sistema CRISPR-Cas (ambos términos se usan indistintamente a lo largo de esta solicitud) no requiere la generación de proteínas personalizadas para dirigirse a secuencias específicas, sino que una molécula de ARN corta puede programar una única enzima Cas para reconocer un ADN diana. Añadir el sistema CRISPR-Cas al repertorio de técnicas de secuenciación del genoma y a los métodos de análisis puede simplificar significativamente la metodología y acelerar la capacidad de catalogar y mapear los factores genéticos asociados con una amplia gama de funciones biológicas y enfermedades. Para utilizar el sistema CRISPR-Cas de manera eficaz para la edición del genoma sin efectos perjudiciales, es fundamental comprender los métodos, sistemas y aparatos para identificar secuencias diana para enzimas Cas o sistemas CRISPR-Cas para secuencias diana de interés y para transmitir los resultados, que son aspectos de la invención reivindicada.

Sumario de la invención

El sistema CRISPR/Cas o el sistema CRISPR-Cas (ambos términos pueden usarse indistintamente a lo largo de esta solicitud) no requiere la generación de proteínas personalizadas para dirigirse a secuencias específicas, sino que una molécula de ARN corta puede programar una única enzima Cas para reconocer un ADN diana, en otras palabras, la enzima Cas puede reclutarse para una diana de ADN específica utilizando dicha molécula de ARN corta. Añadir el sistema CRISPR-Cas al repertorio de técnicas de secuenciación del genoma y a los métodos de análisis puede simplificar significativamente la metodología y acelerar la capacidad de catalogar y mapear los factores genéticos asociados con una amplia gama de funciones biológicas y enfermedades. Para utilizar el sistema CRISPR-Cas de manera eficaz para la edición del genoma sin efectos perjudiciales, es fundamental comprender aspectos de ingeniería y optimización de estas herramientas de ingeniería del genoma, que son aspectos de la invención reivindicada.

En algunos aspectos, la divulgación se refiere a una composición de origen no natural o de ingeniería que comprende una secuencia de polinucleótidos de ARN (ARNchi) quimérico del sistema CRISPR/Cas, en donde la secuencia de polinucleótidos comprende (a) una secuencia guía capaz de hibridar con una secuencia diana en una célula eucariota, (b) una secuencia de apareamiento de tarcr y (c) una secuencia tracr en donde (a), (b) y (c) están dispuestos en una orientación de 5 'a 3', en donde, cuando se transcriben, la secuencia de apareamiento con tarc hibrida con la secuencia tracr y la secuencia guía dirige la unión específica de secuencia de un complejo CRISPR a la secuencia diana, en el que el complejo CRISPR comprende una enzima CRISPR complejada con (1) la secuencia guía que se hibrida con la secuencia diana y (2) la secuencia de apareamiento con tarc que se hibrida con la secuencia tracr,
o
un sistema enzimático CRISPR, en el que el sistema está codificado por un sistema de vectores que comprende uno o más vectores que comprenden I. un primer elemento regulador operativamente unido a una secuencia de polinucleótidos de ARN quimérico (ARNchi) del sistema CRISPR/Cas, en donde la secuencia de polinucleótidos comprende (a) una o más secuencias guía capaces de hibridar con una o más secuencias diana en una célula eucariota, (b) una secuencia de apareamiento con tarc, y (c) una o más secuencias tracr y II. un segundo elemento regulador unido operativamente a una secuencia codificante de enzima que codifica una enzima CRISPR que comprende al menos una o más secuencias de localización nuclear, en donde (a), (b) y (c) están dispuestos en una orientación de 5 'a 3', en donde los componentes I y II están ubicados en el mismo o en diferentes vectores del sistema, en donde, cuando se transcriben, la secuencia de apareamiento con tarc hibrida con la secuencia tracr y la secuencia guía dirige la unión específica de secuencia de un complejo CRISPR a la secuencia diana, en donde el complejo CRISPR comprende la enzima CRISPR complejada con (1) la secuencia guía que se hibrida con la secuencia diana y (2) la secuencia de apareamiento con tarc que se hibrida con la secuencia tracr,
o
un sistema enzimático CRISPR multiplexado, en donde el sistema está codificado por un sistema de vectores que comprende uno o más vectores que comprenden I. un primer elemento regulador operativamente unido a (a) una o más secuencias guía capaces de hibridar con una secuencia diana en una célula, y (b) al menos una o más secuencias de apareamiento con tracr, II. un segundo elemento regulador unido operativamente a una secuencia codificante de enzima que codifica una enzima CRISPR, y III. un tercer elemento regulador unido operativamente a una secuencia tracr, en donde los componentes II y III están ubicados en el mismo o en diferentes vectores del sistema, en donde, cuando se transcriben, la secuencia de apareamiento con tarc hibrida con la secuencia tracr y la secuencia guía dirige la unión específica de secuencia de un complejo CRISPR a la secuencia diana, en donde el complejo CRISPR

comprende la enzima CRISPR complejada con (1) la secuencia guía que se hibrida con la secuencia diana y (2) la secuencia de apareamiento con tarc que se hibrida con la secuencia tracr y en donde en el sistema multiplexado se usan múltiples secuencias guía y una única secuencia tracr.

Sin pretender quedar ligados a teoría alguna, se cree que la secuencia diana debe asociarse con un PAM (motivo adyacente al protoespaciador); es decir, una secuencia corta reconocida por el complejo CRISPR. Este PAM puede considerarse un motivo CRISPR.

Con respecto al sistema o complejo CRISPR analizado en este documento, se hace referencia a la Figura 2. La Figura 2 muestra un sistema CRISPR a modo de ejemplo y un posible mecanismo de acción (A), una adaptación a modo de ejemplo para la expresión en células eucariotas y los resultados de pruebas que evalúan la localización nuclear y la actividad de CRISPR (B-F).

La divulgación proporciona un método para identificar una o más secuencias diana únicas. Las secuencias diana pueden estar en un genoma de un organismo, tal como un genoma de un organismo eucariota. Por consiguiente, a través de un posible enlace específico de secuencia, la secuencia diana puede ser susceptible de ser reconocida por un sistema CRISPR-Cas. (Del mismo modo, la divulgación comprende, por lo tanto, la identificación de uno o más sistemas CRISPR-Cas que identifican una o más secuencias diana únicas). La secuencia diana puede incluir el motivo CRISPR y la secuencia cadena arriba o anterior. El método puede comprender: localizar un motivo CRISPR, por ejemplo, analizar (por ejemplo, comparar) una secuencia para determinar si un motivo CRISPR, por ejemplo, una secuencia PAM, una secuencia corta reconocida por el complejo CRISPR, está presente en la secuencia; analizar (por ejemplo, comparar) la secuencia cadena arriba del motivo CRISPR para determinar si esa secuencia cadena arriba se produce en otro lugar del genoma; seleccionar la secuencia cadena arriba si no ocurre en otro lugar en el genoma, identificando así un sitio diana único. La secuencia cadena arriba del motivo CRISPR puede ser de al menos 10 pb o al menos 11 pb o al menos, 12 pb o al meno,s 13 pb o al menos, 14 pb o al menos, 15 pb o al menos, 16 pb o al menos, 17 pb o al menos, 18 pb o al menos, 19 pb o al menos, 20 pb de longitud, por ejemplo, la secuencia cadena arriba del motivo CRISPR puede ser de aproximadamente 10 pb a aproximadamente 20 pb, por ejemplo, la secuencia cadena arriba es 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 o 30 pb de longitud. El motivo CRISPR puede ser reconocido por una enzima Cas como una enzima Cas9, por ejemplo, una enzima SpCas9. Además, el motivo CRISPR puede ser una secuencia de motivo adyacente al protoespaciador (PAM), por ejemplo, NGG o NAG. Por consiguiente, como los motivos CRISPR o las secuencias PAM pueden ser reconocidas por una enzima Cas *in vitro,* ex vivo o *in vivo,* en el análisis in silico, hay un análisis, por ejemplo, comparación, de la secuencia de interés contra los motivos CRISPR o secuencias PAM para identificar regiones de la secuencia de interés que pueden ser reconocidas por una enzima Cas in vitro, ex vivo o *in vivo.* Cuando ese análisis identifica un motivo CRISPR o una secuencia PAM, el siguiente análisis, por ejemplo, la comparación, es de las secuencias cadena arriba desde el motivo CRISPR o la secuencia PAM, por ejemplo, el análisis de la secuencia 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 o 30 pb de longitud, comenzando por el motivo PAM o CRISPR y extendiéndose cadena arriba desde allí. Ese análisis es para ver si esa secuencia cadena arriba es única, es decir, si la secuencia cadena arriba no parece ocurrir por otra parte en un genoma, puede ser un sitio diana único. La selección para sitios únicos es la misma que la el paso de filtrado: en ambos casos, se filtran todas las secuencias diana con el motivo CRISPR asociado que ocurren más de una vez en el genoma diana.

Los organismos eucariotas de interés pueden incluir, pero sin limitación, *Homo sapiens* (ser humano), *Mus musculus* (ratón)*, Rattus norvegicus* (rata), *Danio rerio* (pez cebra), *Drosophila melanogaster* (mosca de la fruta), *Caenorhabditis elegans* (gusano redondo), *Sus scrofa* (cerdo) y *Bos taurus* (vaca). El organismo eucariota puede seleccionarse del grupo que consiste en *Homo sapiens* (ser humano), *Mus musculus* (ratón)*, Rattus norvegicus* (rata), *Danio rerio* (pez cebra), *Drosophila melanogaster* (mosca de la fruta), *Caenorhabditis elegans* (gusano redondo), *Sus scrofa* (cerdo) y *Bos taurus* (vaca). La divulgación también comprende un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, implementan un método del presente documento para identificar una o más secuencias diana únicas.

La divulgación comprende además un sistema informático para identificar una o más secuencias diana únicas, por ejemplo, en un genoma, tal como un genoma de un organismo eucariota, comprendiendo el sistema: a, una unidad de memoria configurada para recibir y/o almacenar información de secuencia del genoma; y b. uno o más procesadores solos o en combinación programados para realizar un método del presente documento para identificar una o más secuencias diana únicas (por ejemplo, ubicar un motivo CRISPR, analizar una secuencia cadena arriba del motivo CRISPR para determinar si la secuencia ocurre en otro lugar en el genoma, seleccionar la secuencia si no ocurre en otra parte del genoma), para identificar así un sitio diana único y mostrar y/o transmitir una o más secuencias diana únicas. La secuencia diana candidata puede ser una secuencia de ADN. Los desapareamientos pueden ser de ARN del complejo CRISPR y del ADN. En aspectos de la invención, la susceptibilidad de una secuencia diana reconocida por un sistema CRISPR-Cas indica que puede haber una unión estable entre uno o más pares de bases de la secuencia diana y la secuencia guía del sistema CRISPR-Cas para permitir el reconocimiento específico de la secuencia diana por la secuencia guía.
El sistema CRISPR Cas o CRISPR-Cas utiliza una única enzima Cas que puede ser programada por una molécula de ARN corta para reconocer una diana de ADN específica, en otras palabras, la enzima Cas puede reclutarse para una diana de ADN específica utilizando dicha molécula de ARN corta. En determinados aspectos, por ejemplo, cuando no

está mutada o modificada, o cuando está en un estado nativo, la enzima Cas o CRISPR en CRISPR/Cas o el sistema CRISPR-Cas, efectúa un corte en una posición particular; una diana de ADN específica. Por consiguiente, se pueden generar datos- un conjunto de datos de capacitación- relativos al corte por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN en una posición particular para una enzima Cas o CRISPR particular. De manera similar, se pueden generar datos- un conjunto de datos de capacitación- relativos al corte por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN de un desapareamiento particular de la hibridación de ácido nucleico típica (por ejemplo, en lugar de G-C en una posición particular, G-T o G-U o G-A o G-G) para la Cas particular. Al generar tales conjuntos de datos, existe el concepto de frecuencia de corte promedio. La frecuencia con la que una enzima corta una molécula de ácido nucleico, por ejemplo, ADN, es principalmente una función de la longitud de la secuencia a la que es sensible. Por ejemplo, si una enzima tiene una secuencia de reconocimiento de 4 pares de bases, fuera de toda probabilidad, con 4 posiciones, y cada posición tiene potencialmente 4 valores diferentes, hay 44 o 256 posibilidades diferentes para cualquier cadena de 4 bases de largo. Por lo tanto, teóricamente (asumiendo un ADN completamente aleatorio), esta enzima cortará 1 en 256 sitios de 4 pares de bases de largo. Para una enzima que reconoce una secuencia de 6 pares de bases, el cálculo es de 46 o 4096 combinaciones posibles con esta longitud, por lo que tal enzima cortará 1 en 4096 sitios de 6 pares de bases de largo. Por supuesto, dichos cálculos solo tienen en cuenta que cada posición tiene potencialmente 4 valores diferentes y un ADN completamente aleatorio. Sin embargo, el ADN no es completamente aleatorio; por ejemplo, el contenido de G-C de los organismos varía. Por consiguiente, los conjuntos de datos de capacitación en la invención provienen de la observación de cortes por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN en una posición particular para una enzima Cas o CRISPR particular y observando el corte por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN de un desapareamiento particular de la hibridación típica de ácido nucleico para la Cas particular, en un número estadísticamente significativo de experimentos en cuanto a la posición particular, el sistema CRISPR-Cas y la Cas particular, y promediando los resultados observados u obtenidos del mismo. La frecuencia de corte promedio se puede definir como la media de las eficacias de escisión para todos los desapareamientos ARN guía:ADN diana en una ubicación particular.

La divulgación proporciona, adicionalmente, un método para identificar una o más secuencias diana únicas, por ejemplo, en un genoma, tal como un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas (y asimismo, la divulgación también proporciona un método para identificar un sistema CRISPR-Cas susceptible de reconocer una o más secuencias diana únicas), en donde el método comprende: a) determinar la frecuencia de corte promedio en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación en cuanto a esa Cas, b) determinar la frecuencia de corte promedio de un desapareamiento particular (por ejemplo, desapareamiento ARN guía/diana) para la Cas particular a partir del conjunto de datos de capacitación, c) multiplicar la frecuencia de corte promedio en una posición particular por la frecuencia de corte promedio de un desapareamiento particular para obtener un primer producto, d) repetir los pasos a) a c) para obtener un segundo y más productos para cualquier posición particular adicional de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular (u opcionalmente d) repetir los pasos a) a c) para obtener un segundo y más productos para cualquier posición particular adicional de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular), y e) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por la distancia, en pb, entre la primera y la última base de la secuencia diana, por ejemplo, 15-20, tal como 18, y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular(u opcionalmente e) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por la distancia, en pb, entre la primera y la última base de la secuencia diana, por ejemplo, 15-20, tal como 18 y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular), para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. Los pasos (a) y (b) se pueden realizar en cualquier orden. Si no hay más productos que el primer producto, ese primer producto (del paso (c) de multiplicar (a) veces (b)) es el que se utiliza para determinar u obtener la clasificación.

La divulgación también comprende un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) crear un conjunto de datos de capacitación en cuanto a una Cas particular, b) determinar la frecuencia de corte promedio en una posición particular para la Cas particular del conjunto de datos de capacitación, c) determinar la frecuencia de corte promedio de un desajuste particular para la Cas particular a partir del conjunto de datos de capacitación, d) multiplicar la frecuencia de corte promedio en una posición particular por la frecuencia de corte promedio de un desapareamiento particular para obtener un primer producto, e) repetir los pasos b) a d) para obtener un segundo y más productos para cualquier posición particular adicional de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular (u opcionalmente e) repetir los pasos b) a d) para obtener un segundo y más productos para

cualquier posición particular adicional de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento en particular en una posición particular), y f) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por 18 y omitir este paso si no hay un desapareamiento en ninguna posición o si hay solo un desapareamiento particular en una posición particular (u opcionalmente f) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por la distancia, en pb, entre la primera y la última base de la secuencia diana, por ejemplo, 15-20, tal como 18 y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular), para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. Los pasos (a) y (b) se pueden realizar en cualquier orden. Los pasos (a) y (b) se pueden realizar en cualquier orden. Si no hay más productos que el primer producto, ese primer producto (del paso (c) de multiplicar (a) veces (b)) es el que se utiliza para determinar u obtener la clasificación.

La divulgación también comprende un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) determinar la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación para esa Cas, y/o b) determinar la frecuencia de corte promedio de un tipo particular de desapareamiento para la Cas particular del conjunto de datos de capacitación, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. El método puede comprender determinar tanto la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación de esa Cas, y la frecuencia de corte promedio de un tipo particular de desapareamiento para la Cas particular del conjunto de datos de capacitación. Cuando se determinan ambas, el método puede comprender además multiplicar la frecuencia de corte promedio en una posición particular por la frecuencia de corte promedio de un tipo de desapareamiento particular para obtener un primer producto, repetir los pasos de determinación y multiplicación para obtener un segundo y más productos para cualquier posición particular adicional de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular y multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos entre la distancia, en pb, entre la primera y la última base de la secuencia diana y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición en particular, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. La distancia, en pb, entre la primera y la última base de la secuencia diana puede ser 18. El método puede comprender crear un conjunto de capacitación como para una Cas particular. El método puede comprender determinar la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación en cuanto a esa Cas, si hay más de un desapareamiento, repitiendo el paso determinante para determinar la frecuencia de corte para cada desapareamiento y multiplicar las frecuencias de los desapareamientos para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas.

La divulgación comprende adicionalmente un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) determinar la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación para esa Cas y la frecuencia de corte promedio de un tipo particular de desapareamiento para la Cas particular del conjunto de datos de capacitación, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. La divulgación comprende adicionalmente un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) crear un conjunto de datos de capacitación en cuanto a una Cas particular, b) determinar la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para la Cas particular a partir del conjunto de datos de capacitación, y/o c) determinar la frecuencia de corte promedio de un tipo de desapareamiento particular para la Cas particular del conjunto de datos de capacitación, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. La divulgación comprende aún más adicionalmente un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) crear un conjunto de datos de capacitación en cuanto a una Cas particular, b)determinar la frecuencia de corte promedio de desapareamientos ARN guía/diana en una posición particular para la Cas particular a partir del conjunto de datos de capacitación y la frecuencia de corte promedio de un tipo particular de desapareamiento para la Cas particular del conjunto de datos de capacitación, para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. Por consiguiente, en estas realizaciones, en lugar de multiplicar los promedios de frecuencia de corte determinados de forma única para una posición de desapareamiento y tipo de desapareamiento por separado, la invención utiliza promedios que se determinan de forma única, por ejemplo, los promedios de frecuencia de corte para un tipo de desapareamiento particular en una posición particular (por lo tanto, sin multiplicar estos, como parte de preparación del conjunto de formación). Estos métodos se pueden realizar de manera iterativa similar a los pasos de los métodos que incluyen la multiplicación, para la determinación de una o más secuencias diana únicas.

La descripción en ciertos aspectos proporciona un método para seleccionar un complejo CRISPR para dirigir y/o escindir una secuencia de ácido nucleico diana candidata dentro de una célula, que comprende los pasos de: (a) determinar la cantidad, ubicación y naturaleza de los desapareamientos de la secuencia guía del (de los) complejo(s) CRISPR potencial(es) y la secuencia de ácido nucleico diana candidata, (b) determinar la contribución de cada cantidad, ubicación y naturaleza de los desapareamientos en la energía libre de hibridación de la unión entre la secuencia de ácido nucleico diana y la secuencia guía del (de los) complejo(s) CRISPR potencial(es) a partir de un conjunto de datos de capacitación, (c) basándose en el análisis de contribución del paso (b), predecir la escisión en las ubicaciones de los desapareamientos de la secuencia de ácido nucleico diana por los complejos CRISPR potenciales, y (d) seleccionar el complejo CRISPR del (de los) complejo(s) CRISPR potencial(es) en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no en las ubicaciones de los desapareamientos) por el complejo CRISPR. E paso (b) puede realizarse: determinando las contribuciones termodinámicas locales, $\Delta G_{ij}(k)$, entre cada secuencia guía i y la secuencia de ácido nucleico diana j en la posición k, en donde se estima $\Delta G_{ij}(k)$ a partir de un algoritmo de predicción bioquímica y α_k es un peso dependiente de la posición calculado a partir del conjunto de datos de capacitación, estimando los valores de la energía libre eficaz *Zij* usando la relación $p_{ij} \square e^{-\beta Z_{ij}}$, en donde p*ij* mide la frecuencia de corte por la secuencia guía i en la secuencia de ácido nucleico diana j y β es una constante de proporcionalidad positiva, que determina los pesos dependientes de la posición α_k encajando los pares espaciador/diana con la suma en todas las *N* bases de la secuencia guía

$$Z_{ij} = \sum_{k=1}^{N} \alpha_k \Delta G_{ij}(k)$$

y en donde, el paso (c) se realiza determinando los pesos dependientes de la posición a partir de la energía libre $\overline{Z_{est}} = G\bar{\alpha}$eficaz entre cada espaciador y cada diana potencial en el genoma y determinando las frecuencias de corte estimadas entre el espaciador y la diana $p_{est} \propto e^{-\beta Z est}$ para, de este modo, predecir la división. Beta se ajusta implícitamente ajustando los valores de alfa (que son completamente libres de multiplicarse, en el proceso de ajuste, por la constante que sea adecuada para Z = suma (alfa*Delta G).

La divulgación también comprende la creación de un conjunto de datos de capacitación. Un conjunto de datos de capacitación son datos de mediciones de frecuencia de corte, obtenidos para maximizar la cobertura y la redundancia ante posibles tipos y posiciones de desapareamientos. Ventajosamente, hay dos paradigmas experimentales para generar un conjunto de datos de capacitación. En un aspecto, la generación de un conjunto de datos comprende analizar la escisión de Cas, por ejemplo, Cas9, en una diana constante y mutar las secuencias guía. En otro aspecto, la generación de un conjunto de datos comprende analizar la escisión de Cas, por ejemplo, Cas9, utilizando una secuencia guía constante y probar la escisión de múltiples dianas de ADN. Además, el método se puede realizar al menos de dos maneras: *in vivo* (en células, tejido o animal vivo) o *in vitro* (con un ensayo sin células, usando ARN guía y Cas transcritos *in vitro*, por ejemplo, proteína Cas9 administrada por lisado celular completo o proteína purificada). Ventajosamente, el método se realiza analizando la escisión de una diana constante con ARN guía desapareado *in vivo* en líneas celulares. Debido a que el ARN guía puede generarse en las células como un transcrito de un promotor de ARN polimerasa III (por ejemplo, U6) que conduce un oligo de ADN, puede expresarse como un casete de PCR y transfectar el ARN guía directamente (Fig. 24c) junto con Cas9 dirigida por CBh (PX165, Fig. 24c). Co-transfectando Cas9 y un ARN guía con uno o varios desapareamientos con respecto a la diana de ADN constante, se puede evaluar la escisión en un locus endógeno constante mediante un ensayo de nucleasa como el ensayo de nucleasa SURVEYOR o por secuenciación profunda de nueva generación. Estos datos se pueden recopilar para al menos una o varias dianas dentro de unos loci de interés, por ejemplo, al menos 1, al menos 5, al menos 10, al menos 15 o al menos 20 dianas del locus de EMX1humano. De esta manera, se puede generar fácilmente un conjunto de datos de capacitación para cualquier locus de interés. Por consiguiente, hay al menos dos formas de generar un conjunto de datos de capacitación - *in vivo* (en líneas celulares o animales vivos) o *in vitro* (con un ensayo sin células), usando ARN guía y Cas transcritos *in vitro*, por ejemplo, Cas9, proteína administrada por lisado celular completo o proteína purificada). Además, el paradigma experimental puede diferir, p. ej. con secuencias guía mutadas o con una guía constante y una biblioteca de oligos de muchas dianas de ADN. Estos experimentos de direccionamiento también se pueden hacer *in vitro*. La lectura simplemente sería procesar un gel sobre el resultado del ensayo de escisión *in vitro*- los resultados serán fracciones escindidas y no escindidas. Como alternativa o adicionalmente, estas fracciones se pueden aislar en gel y los adaptadores de secuenciación pueden ligarse antes de la secuenciación profunda en estas poblaciones.

La divulgación comprende un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, implementa un método del presente documento. La divulgación comprende además un sistema informático para realizar un método del presente documento. El sistema puede incluir I. una unidad de memoria configurada para recibir y/o almacenar información de secuencia del genoma; y II. uno o más procesadores solos o en combinación programados para llevar a cabo el método del presente documento, por lo que se muestra o transmite la identificación de una o más secuencias diana únicas de manera ventajosa. El organismo eucariota puede seleccionarse del grupo que consiste en *Homo sapiens* (ser humano), *Mus musculus* (ratón)*, Rattus norvegicus* (rata),

Danio rerio (pez cebra), *Drosophila melanogaster* (mosca de la fruta), *Caenorhabditis elegans* (gusano redondo), *Sus scrofa* (cerdo) y *Bos taurus* (vaca). La secuencia diana puede ser una secuencia de ADN y los desapareamientos pueden ser de ARN del complejo CRISPR y el ADN.

La divulgación también conlleva un método para seleccionar un complejo CRISPR para dirigir y/o escindir una secuencia de ácido nucleico diana candidata, por ejemplo, dentro de una célula, que comprende los pasos de: (a) determinar la cantidad, ubicación y naturaleza del (de los) desapareamiento(s) del (de los) complejo(s) CRISPR potencial(es) y la secuencia de ácido nucleico diana candidata, (b) determinar la contribución del (los) desapareamiento(s) en función de la cantidad y la ubicación del (los) desapareamiento(s), (c) basándose en el análisis de contribución del paso (b), predecir la escisión en la ubicación o en las ubicaciones del (de los) desapareamiento(s) y (d) seleccionar el complejo CRISPR del (de los) complejo(s) CRISPR potencial(es) en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no, en la ubicación o en las ubicaciones del(de los) desapareamiento(s) por el complejo CRISPR. La célula puede ser de un organismo eucariota como se analiza en el presente documento. Los pasos determinantes pueden basarse en los resultados o en los conjuntos de datos de capacitación en la invención que provienen de observar el corte por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN en una posición particular para una enzima Cas o CRISPR particular y observando el corte por un sistema CRISPR-Cas en una posición particular en un nucleótido, por ejemplo, secuencia de ADN de un desapareamiento particular de la hibridación típica de ácido nucleico para la Cas particular, en un número estadísticamente significativo de experimentos en cuanto a la posición particular, el sistema CRISPR-Cas y la Cas particular, y promediando los resultados observados u obtenidos del mismo. Por consiguiente, por ejemplo, si el conjunto de datos de capacitación muestra que en una posición particular, el sistema CRISPR-Cas que incluye una Cas particular, es bastante heterogéneo, es decir, puede haber desapareamientos y cortes, la cantidad y la ubicación pueden ser una posición, y la naturaleza del desapareamiento entre el complejo CRISPR y la secuencia de ácido nucleico diana candidata puede no ser importante, por lo que la contribución del desapareamiento al fracaso para cortar/unir puede ser despreciable y la predicción para la escisión puede ser más probable que se produzca esa escisión a que no, a pesar del desapareamiento. Por consiguiente, debe quedar claro que los conjuntos de datos de capacitación no se generan in silico sino que se generan en el laboratorio, por ejemplo, son de estudios *in vitro, ex vivo* y/o *in vivo*. Los resultados del trabajo de laboratorio, por ejemplo, de estudios *in vitro, ex vivo* y/o *in vivo*, se introducen en los sistemas informáticos para realizar los métodos del presente documento.

En los métodos del presente documento, la secuencia diana candidata puede ser una secuencia de ADN, y los desapareamientos pueden ser del ARN del (de los) complejo(s) CRISPR potencial(es) y el ADN. La cantidad de desapareamientos indica el número de desapareamientos en el emparejamiento de bases de ADN:ARN entre el ADN de la secuencia diana y el ARN de la secuencia guía. La ubicación de los desapareamientos indica la ubicación específica a lo largo de la secuencia ocupada por el desapareamiento y, si hay más de un desapareamiento, si los desapareamientos están concatenados u ocurren consecutivamente o si están separados por al menos uno o más restos. En aspectos de la invención, la naturaleza de los desapareamientos indica el tipo de nucleótido implicado en el emparejamiento de bases desaparaeadas. Los pares de bases se aparean de acuerdo con el emparejamiento bases de G-C y A-U de Watson-Crick.

La divulgación implica además un método para predecir la eficacia de la escisión en la secuencia de ácido nucleico diana candidata, por ejemplo, dentro de una diana en una célula, mediante un complejo CRISPR que comprende los pasos de: (a) determinar la cantidad, ubicación y naturaleza del (los) desapareamiento(s) de los complejos CRISPR y la secuencia de ácido nucleico diana candidata, (b) determinar la contribución del (los) desapareamiento(s) en función de la cantidad y la ubicación del (los) desapareamiento(s), y (c) basándose en el análisis de contribución del paso (b), predecir si es más probable que no se produzca la escisión en la ubicación o en las ubicaciones del (de los) desapareamiento(s) y, por lo tanto, predecir la escisión. Al igual que con otros métodos del presente documento, la secuencia diana puede ser una secuencia de ADN y el (los) desapareamiento(s) puede(n) ser de ARN del complejo CRISPR y el ADN. La célula puede ser de un organismo eucariota como se analiza en el presente documento.

La divulgación proporciona además un método para seleccionar una secuencia diana candidata, por ejemplo, dentro de una secuencia de ácido nucleico, por ejemplo, en una célula, para el direccionamiento por un complejo CRISPR, que comprende los pasos de: determinar las contribuciones termodinámicas locales, $\Delta G_{ij}(k)$, entre cada espaciador *i* y diana *j* en la posición *k*, expresar una energía libre eficaz Z_{ij} para cada par espaciador/diana como la suma

$$Z_{ij} = \sum_{k=1}^{N} \alpha_k \Delta G_{ij}(k)$$

en donde ΔG_{ij} *(k)* son contribuciones termodinámicas locales, estimadas a partir de un algoritmo de predicción bioquímica y α_k son los pesos dependientes de la posición, y estimar la energía libre eficaz Z a través de la relación $p_{ij} \sim e^{-\beta Z_{ij}}$ en donde p_{ij} es la frecuencia de corte medida por el espaciador *i* en la diana *j* y β es un ajuste constante positivo en todo el conjunto de datos, y estimar los pesos dependientes de la posición α_k ajustando de tal manera que $G\vec{\alpha} = \vec{Z}$ cada par espaciador-diana(i, j) corresponde a una la fila en la matriz G y cada posición *k* en el emparejamiento

espaciador-diana corresponde a una columna en la misma matriz, y estimar la energía libre eficaz $\overline{E_{est}} = G\vec{\varepsilon}$ entre cada espaciador y cada diana potencial en el genoma usando los valores ajustados α_k, y seleccionar, basándose en los valores de energía libre eficaz calculados, el par espaciador/diana candidato ij de acuerdo con su especificidad y/o eficacia, dadas las frecuencias de corte espaciador-diana estimadas $p_{est} \propto e^{-\beta Zest}$. La célula puede ser de un organismo eucariota como se analiza en el presente documento.

La divulgación también incluye un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, implementa un método para seleccionar un complejo CRISPR para dirigir y/o escindir un ácido nucleico diana candidato, por ejemplo, secuencia dentro de una célula, que comprende los pasos de: (a) determinar la cantidad, ubicación y naturaleza del (de los) desapareamiento(s) del (de los) complejo(s) CRISPR potencial(es) y la secuencia de ácido nucleico diana candidata, (b) determinar la contribución del (los) desapareamiento(s) en función de la cantidad y la ubicación del (los) desapareamiento(s), (c) basándose en el análisis de contribución del paso (b), predecir la escisión en la ubicación o en las ubicaciones del (de los) desapareamiento(s) y (d) seleccionar el complejo CRISPR del (de los) complejo(s) CRISPR potencial(es) en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no, en la ubicación o en las ubicaciones del(de los) desapareamiento(s) por el complejo CRISPR. La célula puede ser de un organismo eucariota como se analiza en el presente documento.

Además, la invención implica sistemas informáticos para seleccionar un complejo CRISPR para dirigir y/o escindir una secuencia de ácido nucleico diana candidata, por ejemplo, dentro de una célula, comprendiendo el sistema: a. una unidad de memoria configurada para recibir y/o almacenar información de secuencia de la secuencia de ácido nucleico diana candidata; y b. uno o más procesadores solos o en combinación programados para (a) determinar la cantidad, ubicación y naturaleza del (de los) desapareamiento(s) del (de los) complejo(s) CRISPR potencial(es) y la secuencia de ácido nucleico diana candidata, (b) determinar la contribución del (de los) desapareamientos en función de la cantidad y la ubicación del (de los) desapareamientos, (c) basándose en el análisis de contribución del paso (b), predecir la escisión en la(s) ubicación (ubicaciones) del (de los) desapareamientos y (d) seleccionar el complejo CRISPR del (de los) complejo(s) CRISPR potencial(es) en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no en las ubicaciones de los desapareamientos por el complejo CRISPR. La célula puede ser de un organismo eucariota como se analiza en el presente documento. El sistema puede mostrar o transmitir la selección.

En aspectos de la invención mencionada en el presente documento, la cantidad de desapareamientos indica el número de desapareamientos en el emparejamiento de bases de ADN:ARN entre el ADN de la secuencia diana y el ARN de la secuencia guía. En aspectos de la invención la ubicación de los desapareamientos indica la ubicación específica a lo largo de la secuencia ocupada por el desapareamiento y, si hay más de un desapareamiento, si los desapareamientos están concatenados u ocurren consecutivamente o si están separados por al menos uno o más restos. En aspectos de la invención, la naturaleza de los desapareamientos indica el tipo de nucleótido implicado en el emparejamiento de bases desaparaeadas. Los pares de bases se aparean de acuerdo con el emparejamiento de bases de G-C y A-U de Watson-Crick.

Por consiguiente, aspectos de la divulgación se relacionan con los métodos y composiciones utilizados para determinar la especificidad de Cas9. En un aspecto, la posición y el número de desapareamientos en el ARN guía se evalúan para la eficacia de escisión. Esta información permite el diseño de secuencias diana que tienen un mínimo de efectos inespecíficos.

La divulgación también comprende un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en el que el método comprende a) determinar la frecuencia de corte promedio de los desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación para esa Cas, y si hay más de un desapareamiento, entonces el paso a) se repite para determinar la frecuencia de corte para cada desapareamiento después de lo cual las frecuencias de los desapareamientos se multiplican para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. La divulgación comprende adicionalmente un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, por lo que la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende a) crear un conjunto de datos de capacitación en cuanto a una Cas particular, b) determinar la frecuencia de corte promedio de los desapareamientos ARN guía/diana en una posición particular para una Cas particular del conjunto de datos de capacitación, y si existe más de un desapareamiento, repetir el paso b) para determinar la frecuencia de corte para cada desapareamiento, a continuación, multiplicar las frecuencias de los desapareamientos para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas. La divulgación también se refiere a los sistemas informáticos y los medios legibles por ordenador que ejecutan estos métodos.

En diversos aspectos, la divulgación implica un sistema informático para seleccionar una secuencia diana candidata dentro de una secuencia de ácido nucleico o para seleccionar una Cas para una secuencia diana candidata, por ejemplo, seleccionar una diana en una célula eucariota para el direccionamiento por un complejo CRISPR.

El sistema informático puede comprender: a. una unidad de memoria configurada para recibir y/o almacenar dicha secuencia de ácido nucleico; y (b) uno o más procesadores solos o en combinación programados para llevarlo a cabo como se analiza en el presente documento. Por ejemplo, programado para: (i) localizar una secuencia de motivos CRISPR (por ejemplo, PAM) dentro de dicha secuencia de ácido nucleico, y (ii) seleccionar una secuencia adyacente a dicha secuencia de motivos CRISPR localizada (por ejemplo, PAM) como la secuencia diana candidata a la que se une el complejo CRISPR. En algunas realizaciones, dicho paso de localización puede comprender identificar una secuencia de motivo CRISPR (por ejemplo, PAM) ubicada a menos de aproximadamente 10000 nucleótidos de distancia de dicha secuencia diana, tal como a menos de aproximadamente 5000, 2500, 1000, 500, 250, 100, 50, 25 o menos nucleótidos de distancia de la secuencia diana. En algunas realizaciones, la secuencia diana candidata tiene una longitud de al menos 10, 15, 20, 25, 30 o más nucleótidos. En algunas realizaciones la secuencia diana candidata tiene una longitud de 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39 o 40 nucleótidos. En algunas realizaciones, el nucleótido en el extremo 3 'de la secuencia diana candidata se ubica no más de aproximadamente 10 nucleótidos cadena arriba de la secuencia del motivo CRISPR (por ejemplo, PAM), tal como no más de 5, 4, 3, 2 o 1 nucleótidos. En algunas realizaciones, la secuencia de ácido nucleico en la célula eucariota es endógena a la célula u organismo, por ejemplo, genoma eucariota. En algunas realizaciones, la secuencia de ácido nucleico en la célula eucariota es exógena a la célula u organismo, por ejemplo, genoma eucariota.

En diversos aspectos, la divulgación proporciona un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, implementa un método descrito en el presente documento, por ejemplo, para seleccionar una secuencia diana candidata dentro de una secuencia de ácido nucleico o seleccionar un CRISPR candidato para una secuencia diana; por ejemplo, una secuencia diana en una célula tal como una célula eucariota para el direccionamiento por un complejo CRISPR. El método puede comprender: (i) localizar una secuencia de motivos CRISPR (por ejemplo, PAM) dentro de dicha secuencia de ácido nucleico, y (ii) seleccionar una secuencia adyacente a dicha secuencia de motivos CRISPR localizada (por ejemplo, PAM) como la secuencia diana candidata a la que se une el complejo CRISPR. En algunas realizaciones, dicho paso de localización puede comprender identificar una secuencia de motivo CRISPR (por ejemplo, PAM) ubicada a menos de aproximadamente 10000 nucleótidos de distancia de dicha secuencia diana, tal como a menos de aproximadamente 5000, 2500, 1000, 500, 250, 100, 50, 25 o menos nucleótidos de distancia de la secuencia diana. En algunas realizaciones, la secuencia diana candidata tiene una longitud de al menos 10, 15, 20, 25, 30 o más nucleótidos. En algunas realizaciones la secuencia diana candidata tiene una longitud de 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39 o 40 nucleótidos. En algunas realizaciones, el nucleótido en el extremo 3 'de la secuencia diana candidata se ubica no más de aproximadamente 10 nucleótidos cadena arriba de la secuencia del motivo CRISPR (por ejemplo, PAM), tal como no más de 5, 4, 3, 2 o 1 nucleótidos. En algunas realizaciones, la secuencia de ácido nucleico en la célula eucariota es endógena a la célula u organismo, por ejemplo, genoma eucariota. En algunas realizaciones, la secuencia de ácido nucleico en la célula eucariota es exógena a la célula u organismo, por ejemplo, genoma eucariota,

Se puede utilizar un sistema informático (o dispositivo digital) para recibir, transmitir, mostrar y/o almacenar resultados, analizar los resultados y/o producir un informe de los resultados y el análisis. Un sistema informático puede entenderse como un aparato lógico que puede leer instrucciones de medios (por ejemplo, software) y/o puerto de red (por ejemplo, de Internet), que puede conectarse opcionalmente a un servidor que tiene medios fijos. Estos sistemas informáticos pueden comprender una o más de una CPU, unidades de disco, dispositivos de entrada tal como teclado y/o ratón, y una pantalla (por ejemplo, un monitor). La comunicación de datos, tal como la transmisión de instrucciones o informes, se puede lograr a través de un medio de comunicación a un servidor en una ubicación local o remota. El medio de comunicación puede incluir cualquier medio de transmisión y/o recepción de datos. Por ejemplo, el medio de comunicación puede ser una conexión de red, una conexión inalámbrica o una conexión a Internet. Dicha conexión puede proporcionar comunicación a través de la World Wide Web. Se prevé que los datos relacionados con la presente invención puedan transmitirse a través de dichas redes o conexiones (o cualquier otro medio adecuado para transmitir información, que incluye, pero sin limitación, el envío de un informe físico, como una impresión) para su recepción y/o para su revisión por un receptor. El receptor puede ser, pero sin limitación, un sistema individual o electrónico (por ejemplo, uno o más ordenadores y/o uno o más servidores). En algunas realizaciones, el sistema informático comprende uno o más procesadores. Los procesadores se pueden asociar con uno o más controladores, unidades de cálculo y/u otras unidades de un sistema informático, o se pueden implantar en el firmware según se desee. Si se implementa en el software, las rutinas se pueden almacenar en cualquier memoria legible por ordenador, como en la memoria RAM, ROM, memoria flash, un disco magnético, un disco láser u otro medio de almacenamiento adecuado. Asimismo, este software se puede suministrar a un dispositivo informático a través de cualquier método de suministro conocido que incluye, por ejemplo, a través de un canal de comunicación tal como una línea telefónica, internet, una conexión inalámbrica, etc., o a través de un medio transportable, tal como un disco legible por ordenador, unidad flash, etc. Los distintos pasos pueden implementarse como varios bloques, operaciones, herramientas, módulos y técnicas que, a su vez, pueden implementarse en hardware, firmware, software o cualquier combinación de hardware, firmware y/o software. Cuando se implementa en hardware, algunos o todos los bloques, operaciones, técnicas, etc. se pueden implementar en, por ejemplo, un circuito integrado personalizado (IC), un circuito integrado específico de la aplicación (ASIC), una matriz lógica programable de campo (FPGA), una matriz lógica programable (PLA), etc.

Se puede usar una arquitectura de base de datos relacional cliente-servidor en realizaciones de la invención, una

arquitectura cliente-servidor es una arquitectura de red en la que cada ordenador o proceso en la red es un cliente o un servidor. Los ordenadores de servidor suelen ser potentes ordenadores dedicados a administrar unidades de disco (servidores de archivos), impresoras (servidores de impresión) o tráfico de red (servidores de red). Los ordenadores cliente incluyen PC (ordenadores personales) o estaciones de trabajo en las que los usuarios ejecutan aplicaciones, así como ejemplos de dispositivos de salida como se describe en el presente documento. Los ordenadores cliente dependen de los ordenadores del servidor para obtener recursos, tales como archivos, dispositivos, e incluso poder de procesamiento. En algunas realizaciones, el ordenador del servidor maneja toda la funcionalidad de la base de datos. El ordenador cliente puede tener un software que maneja toda la administración de datos de interfaz y también puede recibir datos ingresados por los usuarios.

Un medio legible por máquina que comprende un código ejecutable por ordenador puede tomar muchas formas, incluyendo, pero sin limitación, un medio de almacenamiento tangible, un medio de onda portadora o un medio de transmisión físico. Los medios de almacenamiento no volátiles incluyen, por ejemplo, discos ópticos o magnéticos, tal como cualquiera de los dispositivos de almacenamiento en cualquier ordenador o similares, como los que se pueden usar para implementar las bases de datos, etc. mostrados en los dibujos. Los medios de almacenamiento volátiles incluyen memoria dinámica, como la memoria principal de dicha plataforma de ordenador. los medios de transmisión tangibles incluyen cables coaxiales; cable de cobre y fibras ópticas, incluidos los cables que forman un bus dentro de un sistema informático. Los medios de transmisión de onda portadora pueden tomar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o de luz, como las generadas durante las comunicaciones de datos por radiofrecuencia (RF) e infrarrojos (IR). Las formas comunes de medios legibles por ordenador incluyen por ejemplo: un disquete, un disco flexible, disco duro, cinta magnética, cualquier otro medio magnético, un CD-ROM, DVD o DVD-ROM, cualquier otro medio óptico, tarjetas de papel perforadas, cualquier otro medio de almacenamiento físico con patrones de orificios, una RAM, una ROM, una PROM y una EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transmita datos o instrucciones, cables o enlaces que transporten dicha onda portadora, o cualquier otro medio desde el cual un ordenador pueda leer código y/o datos de programación. Muchas de estas formas de medios legibles por ordenador pueden estar implicadas en llevar una o más secuencias de una o más instrucciones a un procesador para su ejecución.

El código ejecutable por ordenador sujeto se puede ejecutar en cualquier dispositivo adecuado que comprenda un procesador, incluido un servidor, un PC o un dispositivo móvil como un teléfono inteligente o una tableta. Cualquier controlador u ordenador incluye opcionalmente un monitor, que puede ser una pantalla de tubo de rayos catódicos ("CRT"), una pantalla plana (por ejemplo, cristal líquido de matriz activa, pantalla, pantalla de cristal líquido, etc.) u otros. Los circuitos de ordenador a menudo se colocan en una caja, que incluye numerosos chips de circuitos integrados, como un microprocesador, memoria, circuitos de interfaz y otros. La caja también incluye opcionalmente una unidad de disco duro, una unidad de disquete, una unidad extraíble de alta capacidad, como un CD-ROM grabable, y otros elementos periféricos comunes. Los dispositivos de entrada, tales como el teclado, el ratón o la pantalla táctil, proporcionan opcionalmente la entrada de un usuario. El ordenador puede incluir el software apropiado para recibir instrucciones del usuario, ya sea en forma de entrada del usuario en un conjunto de campos de parámetros, por ejemplo, en una GUI, o en forma de instrucciones preprogramadas, por ejemplo, preprogramadas para varias diferentes operaciones específicas.

Estas y otras realizaciones se divulgan o son evidentes a partir de, la siguiente descripción detallada.

Breve descripción de los dibujos

La figura 1 muestra un esquema de la nucleasa Cas9 guiada por ARN. La nucleasa Cas9 de *Streptococcus pyogenes* está dirigida al ADN genómico mediante un ARN guía sintético (ARNsg) que consiste en una secuencia guía de 20 nt y un armazón. La base de la secuencia guía empareja con la diana de ADN, directamente cadena arriba de un motivo adyacente al protoespaciador 5'-NGG requerido (PAM; magenta), y Cas9 media una ruptura de doble cadena (DSB) ~ 3 pb cadena arriba del PAM (indicada por triángulo).

La Figura 2A-F muestra un sistema CRISPR a modo de ejemplo y un posible mecanismo de acción (A), una adaptación a modo de ejemplo para la expresión en células eucariotas y los resultados de pruebas que evalúan la localización nuclear y la actividad de CRISPR (B-F).

La Figura 3 muestra un ensayo de representación esquemática realizado para evaluar la especificidad de escisión de Cas9 de *Streptococcus pyogenes*. Los desapareamientos de un solo par de bases entre la secuencia de ARN guía y el ADN diana se mapean contra la eficacia de escisión en %.

La Figura 4 muestra un mapeo de mutaciones en la secuencia PAM a la eficacia de escisión en %.

La Figura 5A-C muestra los histogramas de distancias entre el PAM (NGG) del locus 1 de S. pyogenes SF370 (**Figura 5**A) y el PAM del locus 2 de S.thermophilus LMD9 (NNAGAAW) (**Figura 5**B) adyacentes en el genoma humano; y las distancias para cada PAM por cromosoma (Cr) (**Figura 5C**).

La Figura 6A-C muestra la gráfica de la distribución de distancias entre los motivos NGG y NRG en el genoma humano de una manera "superpuesta".

La Figura 7A-D muestra una representación circular del análisis filogenético que revela cinco familias de Cas9s, incluidos tres grupos de Cas9s grandes (~ 1400 aminoácidos) y dos de Cas9s pequeñas (~ 1100 aminoácidos).

La Figura 8A-F muestra una representación lineal del análisis filogenético que revela cinco familias de Cas9s, incluidos tres grupos de Cas9s grandes (~ 1400 aminoácidos) y dos de Cas9s pequeñas (~ 1100 aminoácidos).

La Figura 9A-G muestra la optimización de la arquitectura de ARN guía para la edición del genoma de mamíferos mediada por SpCas9. (a) Esquema del vector de expresión bicistrónico (PX330) para el ARN guía único impulsado por el promotor U6 y para Cas9 de Streptococcus pyogenes (hSpCas9) optimizado por codón humano impulsado por el promotor CBh utilizados para todos los experimentos posteriores. El ARNsg consiste en una secuencia guía de 20 nt (azul) y un armazón (rojo), truncado en varias posiciones como se indica, (b) ensayo SURVEYOR para las indel mediadas por SpCas9 en los loci de EMX1 y PVALB humanos. Las flechas indican los fragmentos de SURVEYOR esperados (n = 3). (c) Análisis de transferencia Northern para las cuatro arquitecturas de truncamiento de ARNsg, con U1 como control de carga, (d) Tanto el tipo silvestre (ts) como el mutante nickasa (D10A) de SpCas9 promovieron la inserción de un sitio HindIII en el gen EMX1 humano. Se utilizaron oligonucleótidos monocatenarios (ODNss), orientados en dirección sentido o antisentido en relación con la secuencia del genoma, como plantillas de recombinación homólogas (Fig. 68). (e) Esquema del locus SERPINB5 humano. Los ARNsg y PAM están indicados por barras de colores encima de la secuencia; la metilcitosina (Me) está resaltada (rosa) y numerada en relación con el sitio de inicio de la transcripción (TSS, +1). (f) Estado de metilación de SERPINB5 analizado mediante secuenciación con bisulfito de 16 clones. Círculos llenos, CpG metilado; círculos vacíos, CpG no metilado. (g) Eficacia de la modificación por tres ARNsg dirigidos a la región metilada de SERPINB5, ensayada por secuenciación profunda (n = 2). Las barras de error indican intervalos de Wilson.

La Figura 10A-C muestra posición, distribución, número e identidad de los desapareamientos de algunos ARN guía que se pueden usar para generar el conjunto de datos de capacitación (estudio sobre la actividad de Cas9 inespecífica).

La Figura 11A-B muestra posiciones adicionales, distribuciones, números e identidades de los desapareamientos de algunos ARN guía que se pueden usar para generar el conjunto de datos de capacitación (estudio sobre la actividad de Cas9 inespecífica).

La Figura 12A-E muestra la eficacia de escisión del desapareamiento único del ARN guía. a, se seleccionaron múltiples sitios diana del locus EMX1 humano. Las bases individuales en las posiciones 1-19 a lo largo de la secuencia de ARN guía, que complementan la secuencia de ADN diana, se mutaron a cada desapareamiento de ribonucleótidos del ARN guía original (azul "N"). b, Actividad de escisión de Cas9 específica para los ARN guía que contienen mutaciones de base única (azul claro: corte alto, azul oscuro: corte bajo) en relación con el ARN guía específico (gris). c, Mapa de calor de transición de base que representa la actividad de escisión relativa de Cas9 para cada posible par de bases ARN:ADN. Las filas se clasificaron en función de la actividad de escisión en las 10 bases PAM-proximales del ARN guía (de alta a baja). Los niveles medios de escisión se calcularon a través de transiciones de base en las 10 bases PAM proximales (barra derecha) y en todas las transiciones en cada posición (barra inferior). El mapa de calor representa datos agregados de mutación de base única de 15 dianas de EMX1, d, Promedio de eficacia de modificación de locus de Cas9 en dianas con todas las posibles secuencias PAM, e, Histograma de distancias entre las apariciones de PAM 5'-NRG dentro del genoma humano. Las secuencias supuestas se identificaron utilizando tanto la cadena positiva como la negativa de las secuencias cromosómicas humanas.

La Figura 13A-C muestra la eficacia de escisión de Cas9 específica con múltiples desapareamientos del ARN guía y la especificidad de todo el genoma, a, eficacia del direccionamiento de Cas9 con ARN guía que contienen desapareamientos concatenados de 2 (arriba), 3 (medio) o 5 (abajo) bases consecutivas para las dianas 1 y 6 de EMX1. Las filas representan diferentes ARN guía mutados y muestran la identidad de cada mutación de nucleótido (celdas blancas; las celdas grises denotan bases no mutadas), b, Cas9 se dirigió a los ARN guía que contenían 3 (arriba, medio) o (4) (abajo) desapareamientos (celdas blancas) separados por diferentes números de bases no mutadas (celdas grises). c, Actividad de escisión en loci diana de EMX1 diana (barra superior) así como en sitios genómicos candidatos inespecíficos. Los posibles loci inespecíficos contenían 1-3 diferencias de bases individuales (celdas blancas) en comparación con los loci específicos.

La Figura 14A-B muestra que SpCas9 escinde dianas metiladas in vitro, a, Las dianas de plásmidos que contienen dinucleótidos CpG se dejan sin metilar o se metilan *in vitro* por M.SssI. Se indican metil-CpG en la secuencia diana o en PAM, b, Escisión mediante SpCas9 de las dianas 1 y 2 no metiladas o metiladas en lisado de células.

La Figura 15 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma humano. Una lista de sitios únicos para el genoma de ser humano, ratón, rata, pez cebra, mosca de la fruta y C. elegans se identificaron por ordenador y se convirtieron en pistas que se pueden visualizar con el navegador del genoma UCSC. Los sitios únicos se definen como aquellos sitios con secuencias semilla (3'-la mayoría de los 12 nucleótidos de la secuencia de espaciador más la secuencia de PAM NGG) que son únicos en todo el genoma.

La Figura 16 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de S. pyogenes en el genoma del ratón.

La Figura 17 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma de rata.

La Figura 18 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma de pez cebra.

La Figura 19 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma de *D. melanogaster*.

La Figura 20 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma *C. elegans*.

La Figura 21 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma de cerdo.

La Figura 22 muestra una pista del navegador del genoma UCSC para identificar sitios diana únicos de Cas9 de *S. pyogenes* en el genoma de vaca.

La figura 23 muestra el Diseñador CRISPR, una aplicación web para la identificación de sitios diana de Cas9. La mayoría de las regiones diana (como los exones) contienen varias posibles secuencias CRISPR ARNsg + PAM. Para minimizar la escisión inespecífica a través del genoma, un canal de ordenador basado en la web clasifica todos los sitios de ARNsg posibles por su especificidad de genoma predicha y genera los cebadores y oligos necesarios para la construcción de cada CRISPR posible, así como los cebadores (a través de Primer3) para el ensayo de alto rendimiento de la escisión inespecífica potencial en un experimento de secuenciación de nueva generación. Optimización de la elección de ARNsg dentro de la secuencia diana de un usuario: El objetivo es minimizar la actividad total inespecífica en todo el genoma humano. Para cada posible elección de ARNsg, hay una identificación de secuencias inespecíficas (que preceden a los PAM NAG o NGG) en todo el genoma humano que contienen hasta 5 pares de bases desapareadas. La eficacia de escisión en cada secuencia inespecífica se predice utilizando un esquema de ponderación derivado experimentalmente. Cada posible ARNsg se clasifica de acuerdo con su escisión inespecífica total predicha; los ARNsg mejor clasificados representan aquellos que probablemente tengan la mayor escisión específica y la menor escisión inespecífica. Además, se facilitan ventajosamente el diseño de reactivos automatizados para la construcción CRISPR, el diseño de cebadores para el ensayo SURVEYOR específico y el diseño de cebadores para la detección y la cuantificación de alto rendimiento de la escisión inespecífica mediante secuenciación de próxima generación.

La Figura 24A-C muestra la selección de dianas y la preparación de los reactivos, (a) Para Cas9 de S. pyogenes, las dianas de 20 pb (resaltadas en azul) deben ir seguidas de 5'-NGG, que puede aparecer en cualquiera de las cadenas del ADN genómico. (b) Esquema para la cotransfección del plásmido de expresión de Cas9 (PX165) y el casete de expresión de ARNsg dirigido por U6 amplificado por PCR. Utilizando una plantilla de PCR que contiene el promotor U6 y un cebador directo fijo (U6 Dir), el ADN codificante de ARNsg puede adjuntarse al cebador inverso de U6 (U6 Inv) y sintetizarse como un oligo de ADN extendido (oligos de ultrameros de IDT). Téngase en cuenta que la secuencia guía (N azules) en U6 Inv es el complemento inverso de la secuencia diana flanqueante 5'-NGG, (c) Esquema para la clonación sin muescas de los oligos de la secuencia guía en un plásmido que contiene Cas9 y armazón de ARNsg (PX330). Los oligos guía (N azules) contienen salientes para la ligadura en el par de sitios BbsI en PS330, con las orientaciones de las cadenas superior e inferior coincidentes con las de la diana genómica (es decir, el oligo superior es la secuencia de 20 pb que precede a 5'-NGG en el ADN genómico). La digestión de PX330 con BbsI permite el reemplazo de los sitios de restricción de Tipo II (contorno azul) con la inserción directa de oligos hibridados. Vale la pena señalar que se colocó una G adicional antes de la primera base de la secuencia guía. Los solicitantes han descubierto que una G adicional delante de la secuencia guía no afecta negativamente a la eficacia del direccionamiento. En los casos en que la secuencia guía de 20 nt de elección no comienza con guanina, la guanina adicional asegurará que el ARNsg sea transcrito de manera eficaz por el promotor U6, que prefiere una guanina en la primera base de la transcripción.

La Figura 25A-E muestra la especificidad de un solo nucleótido de SpCas9. (a) Esquema del diseño experimental. Se analizaron los ARNsg que llevan todos los posibles desapareamientos de pares de bases únicos (N azules) a lo largo de la secuencia guía para cada sitio diana de EMX1 (el sitio diana 1 se muestra como ejemplo), (b) Representación en el mapa de calor de la eficacia relativa de escisión de SpCas9 por 57 ARNsg de mutación única y 1 ARNsg no mutado, cada uno para cuatro sitios diana deEMX1. Para cada diana de EMX1, las identidades de las sustituciones de un solo par de bases se indican a la izquierda; la secuencia guía original se muestra arriba y está resaltada en el mapa de calor (cuadrados grises). Las eficacias de modificación (que aumentan de blanco a azul oscuro) se normalizan a la secuencia guía original, (c) Mapa de calor para la eficacia relativa de escisión de SpCas9 para cada posible par de bases ARN:ADN, compilado a partir de datos agregados de los ARN guía de un solo desapareamiento para 15 dianas de EMX1. Los niveles medios de escisión se calcularon para las10 bases PAM proximales (barra derecha) y en todas las sustituciones en cada posición (barra inferior); las posiciones en gris no fueron cubiertas por los 469 ARNsg de mutación única y los 15 ARNsg no mutados analizados, d) Frecuencias de indel mediadas por SpCas9 en dianas con todas las posibles secuencias PAM, determinadas utilizando el ensayo de nucleasa SURVEYOR. Se probaron dos sitios diana del locus de EMX1 para cada PAM **(Tabla 4)**. (e) Histograma de distancias entre las apariciones de PAM 5'-NRG dentro del genoma humano. Las dianas supuestas se identificaron utilizando ambas cadenas de secuencias cromosómicas humanas (GRCh37/hgl9).

La Figura 26A-C muestra la especificidad de desapareamiento múltiple de SpCas9. (a) Eficacia de escisión de SpCas9 con ARN guía que contienen a, desapareamientos consecutivos de 2, 3 o 5 bases, o (b, c) desapareamientos múltiples separados por diferentes números de bases no mutadas para las dianas 1, 2, 3 y 6 de *EMX1*. Las filas representan cada ARN guía mutado; las sustituciones de nucleótidos se muestran en las celdas blancas; y las celdas grises denotan bases no mutadas. Todas las frecuencias de indel son absolutas y se analizan mediante secuenciación profunda de 2 réplicas biológicas. Las barras de error indican los intervalos de Wilson (Ejemplo 7, Métodos y materiales)

La Figura 27A-D muestra las frecuencias de indel mediadas por SpCas9 en loci genómicos inespecíficos predichos, (a y b) Los niveles de escisión en loci genómicos inespecíficos supuestos que contienen 2 o 3 desapareamientos individuales (celdas blancas) para la diana 1 y la diana 3 de EMX1se analizan mediante secuenciación profunda. La lista de sitios inespecíficos está ordenada por la posición media de las mutaciones. Los supuestos sitios inespecíficos con mutaciones adicionales no mostraron indeles detectables (Tabla 4). La dosis de Cas9 fue de 3×10^{-10} nmol/célula, con un suministro de ARNsg equimolar. Las barras de error indican intervalos de Wilson, (cyd) Frecuencias de Indel para las dianas 1 y 3 de EMX1 y loci inespecíficos seleccionados (OT) en

función de la dosificación de SpCas9 y ARNsg, normalizadas a la escisión específica a la dosis de transfección más alta (n ::: 2). 400 ng a 10 ng de plásmido de Cas9-ARNsg corresponde a 7,1 x 10^{-10} a 1,8 x 10^{-11} nmol/célula. La especificidad de la escisión se mide como una relación de escisión de específica a inespecífica.

La Figura 28A-B muestra el locus de EMX1 humano con los sitios diana. Esquema del locus de EMX1 humano que muestra la ubicación de 15 sitios de ADN diana, indicado por líneas azules con el PAM correspondiente en magenta.

La Figura 29A-B muestra un análisis adicional de sitios inespecíficos genómicos. Los niveles de escisión en loci inespecíficos genómicos candidatos (celdas blancas) para a, diana 2 de EMX1 y b, diana 6 de EMX1 se analizaron mediante secuenciación profunda. Todas las frecuencias de indel son absolutas y se analizan mediante secuenciación profunda de 2 réplicas biológicas. Las barras de error indican los intervalos de confianza de Wilson

La Figura 30 muestra los rangos de frecuencia de corte pronosticadas y observadas entre las dianas de todo el genoma.

La Figura 31 muestra que el PAM para sp. Aureus Cas9 de Staphylococcus aureus es NNGRR.

La figura 32 muestra un diagrama de flujo de los métodos de localización de la invención.

La Figura 33A-B muestra un diagrama de flujo de los métodos termodinámicos de la invención.

La figura 34 muestra un diagrama de flujo de los métodos de multiplicación de la invención.

La figura 35 muestra un diagrama de bloques esquemático de un sistema informático que se puede utilizar para implementar los métodos descritos en el presente documento.

La Figura 36 muestra un gráfico de dispersión de datos de capacitación y datos de prueba no ajustados.

La Figura 37 muestra un diagrama de dispersión de datos de capacitación y datos de prueba ajustados.

Las figuras del presente documento son solo para fines ilustrativos y no están necesariamente dibujadas a escala.

Descripción detallada de la invención

La divulgación se refiere a la ingeniería y optimización de sistemas, métodos y composiciones utilizados para el control de la expresión génica que involucran el direccionamiento de secuencias, como la perturbación del genoma o la edición de genes, que se relacionan con el sistema CRISPR/Cas y sus componentes (Figs. 1 y 2). En realizaciones ventajosas, la enzima Cas es Cas9.

Los términos “polinucleótido", “nucleótido”, "secuencia de nucleótidos", “ácido nucleico” y “oligonucleótido” se usan indistintamente. Se refieren a una forma polimérica de nucleótidos de cualquier longitud, ya sean desoxirribonucleótidos o ribonucleótidos o análogos de los mismos. Los polinucleótidos pueden tener cualquier estructura tridimensional y pueden realizar cualquier función, conocida o desconocida. Los siguientes son ejemplos no limitantes de polinucleótidos: regiones codificantes o no codificantes de un gen o fragmento de gen, loci (locus) definidos a partir del análisis de enlace, exones, intrones, ARN mensajero (ARNm), ARN de transferencia, ARN ribosómico, ARN de interferencia pequeño (ARNip), ARN de horquilla corta (ARNhc), microARN, (miARN), ribozimas, ADNc, polinucleótidos recombinantes, polinucleótidos ramificados, plásmidos, vectores, ADN aislado de cualquier secuencia, ARN aislado de cualquier secuencia, sondas de ácido nucleico y cebadores. El término también abarca estructuras similares a los ácidos nucleicos con esqueletos sintéticos, véase, por ejemplo, Eckstein, 1991; Baserga et al., 1992; Milligan, 1993; documento WO 97/03211; documento WO 96/39154; Mata, 1997; Strauss-Soukup, 1997; y Samstag, 1996. Un polinucleótido puede comprender uno o más nucleótidos modificados, tales como nucleótidos metilados y análogos de nucleótidos. Si está presente, pueden impartirse modificaciones a la estructura de los nucleótidos antes o después del ensamblaje del polímero. La secuencia de nucleótidos puede estar interrumpida por componentes no nucleotídicos. Un polinucleótido puede modificarse adicionalmente después de la polimerización, tal como mediante conjugación con un componente de marcaje.

Como se usa en el presente documento, la expresión "tipo silvestre" es una expresión de la técnica entendido por personas expertas y significa la forma típica de un organismo, cepa, gen o característica tal como aparece en la naturaleza, a diferencia de las formas mutantes o variantes.

Tal como se usa en el presente documento, debe tomarse el término "variante", para indicar la exhibición de cualidades que tienen un patrón que se desvía de lo que ocurre en la naturaleza.

Las expresiones "que ocurren de forma no natural" o "modificados" se usan indistintamente e indican la participación de la mano del hombre. Los términos, cuando se hace referencia a moléculas de ácido nucleico o polipéptidos, significa que la molécula de ácido nucleico o el polipéptido está al menos sustancialmente libre de al menos otro componente con el que están naturalmente asociados en la naturaleza y como se encuentra en la naturaleza.

"Complementariedad" se refiere a la capacidad de un ácido nucleico para formar enlace(s) de hidrógeno con otra secuencia de ácido nucleico ya sea por Watson-Crick tradicional u otros tipos no tradicionales. Un porcentaje de complementariedad indica el porcentaje de restos en una molécula de ácido nucleico que puede formar enlaces de hidrógeno (por ejemplo, el emparejamiento de bases de Watson-Crick) con una segunda secuencia de ácido nucleico (por ejemplo, 5, 6, 7, 8, 9, 10 de cada 10 son el 50%, del 60 %, del 70 %, del 80 %, el 90 %, y el 100 % complementarios). "Perfectamente complementario" significa que todos los restos contiguos de una secuencia de ácido nucleico se unirán por enlace de hidrógeno con el mismo número de restos contiguos en una segunda secuencia

de ácido nucleico. "Substancialmente complementario" como se usa en el presente documento se refiere a un grado de complementariedad que es al menos el 60%, del 65 %, del 70 %, del 75 %, del 80 %, del 85 %, del 90 %, del 95 %, del 97 %, del 98 %, el 99%, o el 100% en una región de 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50 o más nucleótidos, o se refiere a dos ácidos nucleicos que hibridan en condiciones rigurosas.

Como se usa en el presente documento, Las "condiciones rigurosas" para la hibridación se refieren a las condiciones en las que un ácido nucleico que tiene complementariedad con una secuencia diana hibrida predominantemente con la secuencia diana, y sustancialmente no hibrida con secuencias no diana. Las condiciones rigurosas son generalmente dependientes de la secuencia y varían dependiendo de una serie de factores. En general, cuanto más larga sea la secuencia, mayor será la temperatura a la que la secuencia se hibrida específicamente con su secuencia diana. Ejemplos no limitantes de condiciones rigurosas se describen en detalle en Tijssen(1993), Laboratory Techniques In Biochemistry And Molecular Biology- Hybridization With Nucleic Acid Probes Part L Second Chapter “Overview of principles of hybridization and the strategy of nucleic acid probe assay”, Elsevier, N. Y.

"Hibridación" se refiere a una reacción en la que uno o más polinucleótidos reaccionan para formar un complejo que se estabiliza a través de enlaces de hidrógeno entre las bases de los restos de nucleótidos. El enlace de hidrógeno puede ocurrir por el emparejamiento de bases de Watson Crick, la unión de Hoogstein, o de cualquier otra forma específica de secuencia. El complejo puede comprender dos cadenas que forman una estructura dúplex, tres o más cadenas que forman un complejo de múltiples cadenas, una única cadena de auto-hibridación, o cualquier combinación de éstas. Una reacción de hibridación puede constituir un paso en un proceso más extenso, como el inicio de la PCR, o la escisión de un polinucleótido por una enzima. Una secuencia capaz de hibridar con una secuencia dada se conoce como el "complemento" de la secuencia dada.

Como se usa en el presente documento, la expresión "locus genómico" o "locus" (loci en plural) es la ubicación específica de un gen o secuencia de ADN en un cromosoma. Un "gen" se refiere a tramos de ADN o ARN que codifican un polipéptido o una cadena de ARN que tiene un papel funcional que desempeñar en un organismo y, por lo tanto, es la unidad molecular de la herencia en los organismos vivos. Para los fines de esta invención, se puede considerar que los genes incluyen regiones que regulan la producción del producto génico, ya sea que tales secuencias reguladoras sean o no adyacentes a las secuencias codificadas y/o transcritas. Por consiguiente, un gen incluye, pero no se limita necesariamente a, secuencias promotoras, terminadores, secuencias reguladoras de la traducción, como los sitios de unión al ribosoma y los sitios internos de entrada al ribosoma, potenciadores, silenciadores, aisladores, elementos de contorno, orígenes de replicación, sitios de unión de matriz y regiones de control de locus.

Como se usa en el presente documento, "expresión de un locus genómico" o "expresión génica" es el proceso mediante el cual la información de un gen se utiliza en la síntesis de un producto génico funcional. Los productos de la expresión génica son a menudo proteínas, pero en los genes que no codifican proteínas, como los genes de ARNr o los genes de ARNt, el producto es un ARN funcional. El proceso de expresión génica se utiliza por todas las formas de vida conocidas- eucariotas (incluidos los organismos multicelulares), procariotas (bacterias y arqueas) y virus para generar productos funcionales para sobrevivir. Como se usa en el presente documento, "expresión" de un gen o ácido nucleico abarca no solo la expresión génica celular, sino también la transcripción y traducción de ácido(s) nucleico(s) en sistemas de clonación y en cualquier otro contexto. Como se usa en el presente documento, "expresión" también se refiere al proceso por el cual un polinucleótido se transcribe a partir de una plantilla de ADN (tal como en un transcrito de ARNm o de otro ARN) y/o el proceso por el cual un ARNm transcrito se traduce posteriormente en péptidos, polipéptidos o proteínas. Las transcripciones y los polipéptidos codificados pueden denominarse comúnmente "producto genético". Si el polinucleótido deriva de ADN génico, la expresión puede incluir el corte y empalme del ARNm en una célula eucariota.

Los términos “polipéptido”, “péptido” y "proteína" se utilizan indistintamente en el presente documento para referirse a polímeros de aminoácidos de cualquier longitud. El polímero puede ser lineal o ramificado, puede comprender aminoácidos modificados y puede estar interrumpido por no aminoácidos. Los términos también abarcan un polímero de aminoácido que se ha modificado; por ejemplo, formación de enlaces disulfuro, glicosilación, lipidación, acetilación, fosforilación o cualquier otra manipulación, tal como conjugación con un componente de marcaje. Como se usa en el presente documento, el término "aminoácido" incluye aminoácidos naturales y/o no naturales o sintéticos, que incluyen glicina y los isómeros ópticos tanto D como L, y análogos de aminoácidos y peptidométicos.

Como se usa en el presente documento, el término "dominio" o "dominio de proteína" se refiere a una parte de una secuencia de proteína que puede existir y funcionar independientemente del resto de la cadena de proteína.

Como se describe en aspectos de la invención, la identidad de secuencia está relacionada con la homología de secuencia. Las comparaciones de homología pueden realizarse a ojo, o más generalmente, con la ayuda de programas de comparación de secuencias fácilmente disponibles. Estos programas informáticos disponibles comercialmente pueden calcular el porcentaje (%) de homología entre dos o más secuencias y también pueden calcular la identidad de secuencia compartida por dos o más secuencias de aminoácidos o de ácidos nucleicos. En algunas realizaciones preferentes, la región de protección con caperuza de los dTALE descritos en el presente documento tiene secuencias que son al menos el 95% idénticas o comparten identidad con las secuencias de aminoácidos de la región de protección con caperuza que se proporcionan en el presente documento.

Las homologías de secuencia pueden generarse por cualquiera de una serie de programas informáticos conocidos en la técnica, por ejemplo, BLAST o FASTA, etc. Un programa informático adecuado para llevar a cabo dicho alineamiento es el paquete GCG Wisconsin Bestfit (Universidad de Wisconsin, EE. UU.; Devereux *et al.,* 1984, Nucleic Acids Research 12:387). Los ejemplos de otro software que puede realizar comparaciones de secuencias incluyen, pero sin limitación, el paquete BLAST (véase Ausubel et al., 1999 ibid - Capítulo 18), FASTA (Atschul et al., 1990, J. Mol. *Biol*., 403-410) y el conjunto de herramientas de comparación GENEWORKS. Tanto BLAST como FASTA están disponibles para búsquedas fuera de línea y en línea (véase Ausubel et al., 1999 ibid, páginas 7-58 a 7-60). Sin embargo, se prefiere utilizar el programa GCG Bestfit. El % de homología se puede calcular sobre secuencias contiguas, es decir, una secuencia se alinea con la otra secuencia y cada aminoácido o nucleótido en una secuencia se compara directamente con el correspondiente aminoácido o nucleótido en la otra secuencia, un resto cada vez. Esto se llama un alineamiento "sin hueco". Normalmente, tales alineamientos sin huecos se realizan solo en un número relativamente corto de restos. Aunque este es un método muy simple y consistente, no tiene en cuenta que, por ejemplo, en un par de secuencias por lo demás idénticas, una inserción o deleción puede hacer que los siguientes restos de aminoácidos queden fuera de alineamiento, lo que podría generar una gran reducción en el % de homología cuando se produce el alineamiento global. Por consiguiente, la mayoría de los métodos de comparación de secuencias están diseñados para producir alineamientos óptimos que tengan en cuenta posibles inserciones y deleciones sin penalizar indebidamente la homología general o la puntuación de identidad. Esto se logra al insertar "huecos" en el alineamiento de la secuencia para tratar de maximizar la homología o identidad local. Sin embargo, estos métodos más complejos asignan "penalizaciones de hueco" a cada hueco que se produce en el alineamiento, de modo que, para el mismo número de aminoácidos idénticos, un alineamiento de secuencia con la menor cantidad de huecos posible -reflejando una mayor relación entre las dos secuencias comparadas- puede lograr una puntuación más alta que una con muchos huecos. Los “costes de huecos de afinidad” se utilizan normalmente y cargan un coste relativamente alto por la existencia de un hueco y una penalización menor por cada resto posterior en el hueco. Este es el sistema de puntuación de huecos más utilizado. Las altas penalizaciones por huecos pueden, por supuesto, producir alineamientos optimizados con menos huecos. La mayoría de los programas de alineamiento permiten modificar las penalizaciones por huecos. Sin embargo, se prefiere usar los valores predeterminados cuando se usa dicho software para comparaciones de secuencias. Por ejemplo, cuando se utiliza el paquete GCG Wisconsin Bestfit, la penalización por hueco predeterminada para las secuencias de aminoácidos es -12 para un hueco y -4 para cada extensión. El cálculo del % máximo de homología, por lo tanto, requiere primero la producción de un alineamiento óptimo, teniendo en cuenta las penalizaciones por huecos. Un programa informático adecuado para llevar a cabo dicho alineamiento es el paquete GCG Wisconsin Bestfit (Devereux et al., 1984 Nuc. Acids Research 12 p387). Los ejemplos de otro software que puede realizar comparaciones de secuencias incluyen, pero sin limitación, el paquete BLAST (véase Ausubel et al., 1999 Short Protocols in Molecular Biology, 4ª Ed. - Capítulo 18), FASTA (Altschul et al., 1990 J. Mol, *Biol.* 403-410) y el conjunto de herramientas de comparación GENE GENEWORKS. Tanto BLAST como FASTA están disponibles para búsquedas fuera de línea y en línea (véase Ausubel et al., 1999, Short Protocols in Molecular Biology, páginas 7-58 a 7-60). Sin embargo, para algunas aplicaciones, se prefiere utilizar el programa GCG Bestfit. También está disponible una nueva herramienta, llamada Secuencias BLAST 2 para comparar secuencias de proteínas y nucleótidos (véase FEMS Microbiol Lett). 1999 174(2): 247-50; FEMS Microbiol Lett. 1999 177(1): 187-8 y el sitio web de el National Center for Biotechnology information en el sitio web de el National Institutes for Health). Aunque el % de homología final se puede medir en términos de identidad, el proceso de alineamiento en sí no suele basarse en una comparación de pares de todo o nada. En su lugar, generalmente se utiliza una matriz de puntuación de similitud a escala que asigna puntuaciones a cada comparación por pares basada en la similitud química o en la distancia evolutiva. Un ejemplo de una matriz de este tipo utilizada comúnmente es la matriz BLOSUM62, la matriz predeterminada para el conjunto de programas BLAST. Los programas GCG Wisconsin generalmente usan los valores públicos predeterminados o una tabla de comparación de símbolos personalizada, si se suministran (véase el manual del usuario para obtener más detalles). Para algunas aplicaciones, se prefiere utilizar los valores públicos predeterminados para el paquete GCG, o en el caso de otro software, la matriz predeterminada, tal como BLOSUM62.

Como alternativa, el porcentaje de homologías se puede calcular utilizando la función de alineamiento múltiple en DNASISTM (Hitachi Software), basada en un algoritmo, análogo a CLUSTAL (Higgins DG y Sharp PM (1988), Gene 73(1), 237-244). Una vez que el software ha producido una alineación óptima, es posible calcular el % de homología, preferentemente el % de identidad de secuencia. El software normalmente hace esto como parte de la comparación de secuencias y genera un resultado numérico.

Las secuencias también pueden tener deleciones, inserciones o sustituciones de restos de aminoácidos que producen un cambio silencioso y dan como resultado una sustancia funcionalmente equivalente. Se pueden hacer sustituciones de aminoácidos deliberadas sobre la base de la similitud en las propiedades de los aminoácidos (como la polaridad, carga, solubilidad, hidrofobicidad, hidrofilicidad, y/o la naturaleza anfipática de los restos) y, por lo tanto, es útil agrupar los aminoácidos juntos en grupos funcionales. Los aminoácidos pueden agruparse basándose solamente en las propiedades de sus cadenas laterales. Sin embargo, es más útil incluir también datos de mutaciones. Es probable que los conjuntos de aminoácidos derivados de este modo, se conserven por razones estructurales. Estos conjuntos se pueden describir en forma de un diagrama de Venn (Livingstone C.D. y Barton G.J. (1993) “Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation” Comput. Appl. Biosci. 9: 745-756) (Taylor W.R. (1986) “The classification of amino acid conservation” J. Theor. *Biol.* 119; 205-218). Se pueden hacer sustituciones conservativas, por ejemplo, de acuerdo con la tabla a continuación que describe un diagrama de Venn

generalmente aceptado de agrupación de aminoácidos.

Conjunto		Subconjunto	
Hidrófobo	F W Y H K M I L V A G C	Aromático	F W Y H
		Alifático	I L V
Polar	W Y H K R E D C S T N Q	Cargado	H K R E D
		Cargado positivamente	H K R
		Cargado negativamente	E D
Pequeño	V C A G S P T N D	Minúsculo	A G S

Las realizaciones de la divulgación incluyen secuencias (tanto polinucleótidos como polipéptidos) que pueden comprender una sustitución homóloga (tanto la sustitución como el reemplazamiento se utilizan en el presente documento para indicar el intercambio de un resto de aminoácido o nucleótido existente, con un resto o nucleótido alternativo) que puede ocurrir, es decir, sustitución similar-por-similar en el caso de aminoácidos tales como básico por básico, ácido por ácido, polar por polar, etc. También puede producirse una sustitución no homóloga, es decir, de una clase de resto a otra o, implicando alternativamente, la inclusión de aminoácidos no naturales tal como la ornitina (denominada en lo sucesivo, Z), ornitina del ácido diaminobutírico (denominada en lo sucesivo, B), ornitina de norleucina (en lo sucesivo denominada O), pirilalanina, tienilalanina, naftilalanina y fenilglicina.

Las secuencias de aminoácidos variantes pueden incluir grupos espaciadores adecuados que pueden insertarse entre dos restos de aminoácidos cualquiera de la secuencia, incluyendo grupos alquilo tales como grupos metilo, etilo o propilo además de espaciadores de aminoácidos tales como restos de glicina o β-alanina. Una forma adicional de variación, que implica la presencia de uno o más restos de aminoácidos en forma peptoide, puede ser bien entendida por los expertos en la materia. Para disipar cualquier duda, "la forma peptoide" se utiliza para referirse a los restos de aminoácidos variantes en los que el grupo sustituyente carbono α está en el átomo de nitrógeno del resto en lugar del carbono α. Los procesos para preparar péptidos en la forma peptoide son conocidos en la técnica, por ejemplo Simon RJ et al., PNAS (1992) 89(20), 9367-9371 y Horwell DC, Trends Biotechnol. (1995) 13(4), 132-134.

La práctica de la presente invención emplea, a menos que se indique otra cosa, técnicas convencionales de inmunología, bioquímica, química, biología molecular, microbiología, biología celular, genómica y ADN recombinante, que están dentro de las habilidades en la técnica. Véanse Sambrook, Fritsch y Maniatis, MOLECULAR CLONING: A LABORATORY MANUAL, 2ª edición (1989); CURRENT PROTOCOLS IN MOLECULAR BIOLOGY" (F, M. Ausubel, et al. eds., (1987)); the series METHODS IN ENZYMOLOGY (Academic Press, Inc.): PCR 2: A PRACTICAL APPROACH (MJ. MacPherson, B.D. Hames and G.R. Taylor eds. (1995)), Harlow y Lane, ed. (1988) BC, A LABORATORY MANUAL, and ANIMAL CELL CULTURE (R.I. Freshney, ed. (1987)).

En un aspecto, la divulgación proporciona vectores que se utilizan en la ingeniería y optimización de los sistemas CRISPR/Cas. Como se usa en el presente documento, un "vector" es una herramienta que permite o facilita la transferencia de una entidad de un entorno a otro. Es un replicón, tal como un plásmido, fago o cósmido, en el que se puede insertar otro segmento de ADN para lograr la replicación del segmento insertado. En general, un vector es capaz de replicarse cuando está asociado con los elementos de control apropiados. En general, el término "vector" se refiere a una molécula de ácido nucleico capaz de transportar otro ácido nucleico al que se ha unido. Los vectores incluyen, pero sin limitación, moléculas de ácido nucleico que son monocatenarias, bicatenarias o parcialmente bicatenarias; moléculas de ácido nucleico que comprenden uno o más extremos libres, sin extremos libres (por ejemplo, circulares); moléculas de ácido nucleico que comprenden ADN, ARN, o ambos; y otras variedades de polinucleótidos conocidas en la técnica. Un tipo de vector es un "plásmido", que se refiere a un bucle circular de ADN de doble cadena en el que se pueden insertar segmentos de ADN adicionales, tal como por técnicas estándar de clonación molecular. Otro tipo de vector es un vector viral, en donde las secuencias de ADN o ARN derivadas de virus están presentes en el vector para empaquetar en un virus (por ejemplo, retrovirus, retrovirus defectuosos para la replicación, adenovirus, adenovirus defectuosos para la replicación y virus adenoasociados). Los vectores virales también incluyen polinucleótidos transportados por un virus para la transfección en una célula hospedadora. Determinados vectores son capaces de replicarse de manera autónoma en una célula hospedadora en la que se introducen (por ejemplo vectores bacterianos que tienen un origen de replicación bacteriano y vectores episómicos de mamíferos). Otros vectores (por ejemplo, vectores no episómicos de mamífero) se integran en el genoma de una célula hospedadora tras su introducción en la célula hospedadora y, de este modo, se replican junto con el genoma del hospedador. Además, determinados vectores son capaces de dirigir la expresión de genes a los que están unidos operativamente. Tales vectores se denominan en el presente documento "vectores de expresión". Los vectores de expresión comunes de utilidad en las técnicas de ADN recombinante están a menudo en forma de plásmidos. Los vectores de expresión recombinantes pueden comprender un ácido nucleico en una forma adecuada para la expresión del ácido nucleico en una célula hospedadora, lo que significa que los vectores de expresión recombinantes incluyen uno o más elementos reguladores, que pueden seleccionarse basándose en las células hospedadoras que se van a usar para la expresión, que está unido operativamente a la secuencia de ácido nucleico que se va a expresar. Dentro de un vector de expresión recombinante, "operativamente unido" pretende significar que la secuencia de nucleótidos de interés se une al (a los) elemento(s) regulador(es) de una manera que permite la expresión de la secuencia de

nucleótidos (por ejemplo, en un sistema de transcripción/traducción *in vitro)* o en una célula hospedadora cuando se introduce el vector en la célula hospedadora). Con respecto a los métodos de recombinación y clonación, se hace mención de la solicitud de patente de Estados Unidos 10/815.730.

Los aspectos de la divulgación pueden relacionarse con vectores bicistrónicos para ARN quimérico y Cas9. Cas9 está dirigida por el promotor CBh y el ARN quimérico está dirigido por un promotor U6. El ARN guía quimérico consiste en una secuencia guía de 20 pb (Ns) unida a la secuencia tracr (que se extiende desde la primera "U" de la cadena inferior hasta el final de la transcripción), que se trunca en varias posiciones como se indica. Las secuencias guía y tracr se separan mediante la secuencia de apareamiento con tarc GUUUUAGAGCUA seguida por la secuencia de bucle GAAA. Los resultados de los ensayos SURVEYOR para los indeles mediados por Cas9 en los loci EMX1 y PVALB humanos se ilustran en las Figuras 16b y 16c, respectivamente. Las flechas indican los fragmentos de SURVEYOR esperados. Los ARNchi están indicados por su designación "+ n", y ARNcr se refiere a un ARN híbrido donde las secuencias guía y tracr se expresan como transcritos separados. A lo largo de esta solicitud, El ARN quimérico (ARNchi) también se puede llamar guía simple o ARN guía sintético (ARNsg).

La expresión "elemento regulador" pretende incluir promotores, potenciadores, sitios internos de entrada ribosómica (IRES) y otros elementos de control de la expresión (por ejemplo, señales de terminación de la transcripción, tales como señales de poliadenilación y secuencias de poli-U). Dichos elementos reguladores se describen, por ejemplo, en Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990). Los elementos reguladores incluyen aquellos que dirigen la expresión constitutiva de una secuencia de nucleótidos en muchos tipos de células hospedadoras y aquellos que dirigen la expresión de la secuencia de nucleótidos solamente en ciertas células hospedadoras (por ejemplo, secuencias reguladoras específicas de tejido). Un promotor específico de tejido puede dirigir la expresión principalmente en un tejido de interés deseado, tal como músculo, neurona, hueso, piel, sangre, órganos específicos (por ejemplo, hígado, páncreas) o tipos de células particulares (por ejemplo, linfocitos). Los elementos reguladores también pueden dirigir la expresión de una manera temporal dependiente, tal como en una manera dependiente del ciclo celular o dependiente de la etapa de desarrollo, que también puede ser o no específica de tejido o de tipo de célula. En algunas realizaciones, un vector comprende uno o más promotores pol III (por ejemplo, 1, 2, 3, 4, 5 o más promotores pol IIII), uno o más promotores pol II (por ejemplo, 1, 2, 3, 4, 5 o más promotores pol II), uno o más promotores pol I (por ejemplo, 1, 2, 3, 4, 5, o más promotores pol I), o combinaciones de los mismos. Los ejemplos de promotores pol III incluyen, pero sin limitación, promotores U6 y H1. Los ejemplos de promotores pol II incluyen, pero sin limitación, el promotor LTR del virus del sarcoma de Rous retroviral (RSV), (opcionalmente, con el potenciador del RSV), el promotor del citomegalovirus (CMV) (opcionalmente, con el potenciador del CMV) [véase, por ejemplo, Boshart et al., Cell, 41:521-530 (1985)], el promotor de SV40, el promotor de dihidrofolato reductasa, el promotor de a β-actina, el promotor de fosfoglicerol quinasa (PGK) y el promotor de EFla. El término "elemento regulador" también abarca elementos potenciadores, tales como WPRE; potenciadores de CMV; el segmento R-U5 'en LTR de HTLV-I (Mol. Cell. *Biol*., Vol. 8(1), p. 466-472, 1988); potenciador de SV40; y la secuencia de intrones entre los exones 2 y 3 de la β-globina de conejo (Proe, Natl, Acad. Sci. USA., Vol. 78(3), p. 1527-31, 1981). Los expertos en la materia apreciarán que el diseño del vector de expresión puede depender de factores tales como la elección de la célula hospedadora a transformar, el nivel de expresión deseado, etc. Se puede introducir un vector en las células hospedadoras para producir transcritos, proteínas o péptidos, incluyendo proteínas o péptidos de fusión, codificados por los ácidos nucleicos como se describe en el presente documento (por ejemplo, transcritos de repeticiones palindrómicas cortas agrupadas y regularmente interespaciadas (CRISPR), proteínas, enzimas, formas mutantes de los mismos, proteínas de fusión de los mismos, etc.). Con respecto a las secuencias reguladoras, se hace mención de la solicitud de patente de EE.UU. 10/491.026 - Con respecto a los promotores, se hace mención de la publicación PCT WO 2011/028929 y la solicitud de EE.UU. 12/511.940.

Los vectores pueden diseñarse para la expresión de transcritos de CRISPR (por ejemplo, transcritos de ácido nucleico, proteínas o enzimas) en células procariotas o eucariotas. Por ejemplo, Los transcritos de CRISPR pueden expresarse en células bacterianas como *Escherichia coli,* en células de insecto (usando vectores de expresión de baculovirus), células de levadura o células de mamífero. Las células hospedadoras adecuadas se analizan más a fondo en Goeddel, GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990). Como alternativa, el vector de expresión recombinante se puede transcribir y traducir *in vitro,* por ejemplo, utilizando secuencias reguladoras del promotor T7 y polimerasa T7. Los vectores pueden introducirse y propagarse en un procariota o en una célula procariota. En algunas realizaciones, un procariota se usa para amplificar copias de un vector para ser introducido en una célula eucariota o como un vector intermedio en la producción de un vector para ser introducido en una célula eucariota (por ejemplo, amplificar un plásmido como parte de un sistema de empaquetamiento de vector viral). En algunas realizaciones, un procariota se usa para amplificar copias de un vector y expresar uno o más ácidos nucleicos, tal como para proporcionar una fuente de una o más proteínas para el suministro a una célula hospedadora u organismo hospedador. La expresión de proteínas en procariotas se realiza con mayor frecuencia en Escherichia coli con vectores que contienen promotores constitutivos o inducibles que dirigen la expresión de proteínas de fusión o no de fusión. Los vectores de fusión agregan una cantidad de aminoácidos a una proteína codificada en los mismos, tal como al extremo amino de la proteína recombinante. Dichos vectores de fusión pueden servir para uno o más propósitos, tales como: (i) para aumentar la expresión de proteína recombinante; (ii) para aumentar la solubilidad de la proteína recombinante; y (iii) para ayudar en la purificación de la proteína recombinante actuando como ligando en la purificación por afinidad. A menudo, en los vectores de expresión de fusión, se introduce un sitio de escisión proteolítica en la unión del resto de fusión y la proteína recombinante para permitir la

separación de la proteína recombinante del resto de fusión después de la purificación de la proteína de fusión. Dichas enzimas, y sus secuencias de reconocimiento afines, incluyen el factor Xa, la trombina y la enteroquinasa. Ejemplos de vectores de expresión de fusión incluyen pGEX (Pharmacia Biotech Inc; Smith y Johnson, 1988. Gene 67: 31-40), pMAL (New England Biolabs, Beverly, Mass.) y pRIT5 (Pharmacia, Piscataway, N.J) que fusionan glutatión S-transferasa (GST), proteína de unión a maltosa E o proteína A, respectivamente, a la proteína recombinante diana. Ejemplos de vectores de expresión de E. coli inducibles no de fusión adecuados incluyen pTrc (Amrann et al., (1988) Gene 69: 301-315) y pET 11d (Stndier et al., GENE EXPRESSION TECHNOLOGY: METHODS IN ENZYMOLOGY 185, Academic Press, San Diego, Calif. (1990) 60-89). En algunas realizaciones, el vector es un vector de expresión de levadura. Ejemplos de vectores para la expresión en levaduras Saccharomyces cerivisae incluyen pYepSec1 (Baldari, et al., 1987. EMBO J. 6: 229-234), pMFa (Kuijan and Herskowitz, 1982. Cell 30: 933-943), pJRY88 (Schultz et al., 1987. Gene 54: 113-123), pYES2 (Invitrogen Corporation, San Diego, Calif.), y picZ (InVitrogen Corp, San Diego, Calif). En algunas realizaciones, un vector dirige la expresión de proteínas en células de insecto utilizando vectores de expresión de baculovirus. Los vectores de baculovirus disponibles para la expresión de proteínas en células de insecto cultivadas (por ejemplo, células SF9) incluyen la serie pAc (Smith, et al., 1983. Mol. Cell. *Biol.* 3: 2156-2165) y la series pVL (Lucklow y Summers, 1989. Virology 170: 31-39). En algunas realizaciones, un vector es capaz de dirigir la expresión de una o más secuencias en células de mamíferos usando un vector de expresión de mamíferos. Los ejemplos de vectores de expresión de mamífero incluyen pCDM8 (Seed, 1987. Nature 329: 840) y pMT2PC (Kaufman, et al., 1987. EMBO J. 6: 187-195). Cuando se usan en células de mamíferos, las funciones de control del vector de expresión son proporcionadas normalmente por uno o más elementos reguladores. Por ejemplo, los promotores usados comúnmente derivan de polioma, adenovirus 2, citomegalovirus, virus de simio 40 y otros descritos en el presente documento y conocidos en la técnica. Para otros sistemas de expresión adecuados tanto para células procariotas como para eucariotas véase, por ejemplo, capítulos 16 y 17 de Sambrook, et al., MOLECULAR CLONING: A LABORATORY MANUAL,. 2ª ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989.

En algunas realizaciones, el vector de expresión de mamíferos recombinante es capaz de dirigir la expresión del ácido nucleico preferentemente en un tipo de célula particular (por ejemplo, se usan elementos reguladores específicos de tejido para expresar el ácido nucleico). Los elementos reguladores específicos de tejido son conocidos en la técnica. Ejemplos no limitantes de promotores específicos de tejido adecuados incluyen el promotor de albúmina (específico de hígado; Pinkert, et al., 1987. Genes Dev. 1: 268-277), promotores específicos linfoides (Calame y Eaton, 1988. Adv. Immunol. 43: 235-275), en particular promotores de receptores de linfocitos T (Winoto y Baltimore, 1989). EMBO J. 8: 729-733) e inmunoglobulinas (Baneiji, et al., 1983. Cell 33: 729-740; Queen y Baltimore, 1983. Cell 33: 741-748), promotores específicos de neuronas (por ejemplo, el promotor de neurofilamento; Byrne y Ruddle, 1989. Proc. Natl. Acad. Sci. USA 86: 5473-5477), promotores específicos de páncreas (Edlund, et al., 1985. Science 230: 912-916), y promotores específicos de glándula mamaria (por ejemplo, promotor de suero de leche): EE.UU., Pat. N.º 4.873.316 y Publicación de Solicitud Europea N.º 264.166). También se incluyen los promotores regulados por el desarrollo, por ejemplo, los promotores de caja murina (Kessel y Gruss, 1990. Science 249: 374-379) y el promotor de la α-fetoproteína (Campes y Tilghman, 1989. Genes Dev. 3: 537-546). Con respecto a estos vectores procariotas y eucariotas, se hace mención de la patente de EE.UU. 6.750.059. Otras realizaciones de la invención pueden relacionarse con el uso de vectores virales, con respecto a los cuales se hace mención de la solicitud de patente estadounidense 13/092.085. Los elementos reguladores específicos de tejido son conocidos en la técnica y en este sentido, se hace mención de la patente de EE.UU. 7.776.321

En algunas realizaciones, un elemento regulador está operativamente unido a uno o más elementos de un sistema CRISPR para dirigir la expresión de uno o más elementos del sistema CRISPR. En general, las CRISPR (Repeticiones Palindrómicas Cortas Agrupadas y Regularmente Interespaciadas), también conocidas como SPIDR (Repeticiones Directas Intercaladas por Espaciadores), constituyen una familia de loci de ADN que generalmente son específicas de una especie bacteriana en particular. El locus CRISPR comprende una clase distinta de repeticiones de secuencia corta (SSR) intercaladas que se reconocieron en E. coli (Ishino et al., J. Bacteriol., 169:5429-5433 [1987]; y Nakata et al., J. Bacteriol., 171:3553-3556 [1989]) y genes asociados. Se han identificado SSR intercaladas similares en Haloferax mediterranei, Streptococcus pyogenes, Anabaena y Mycobacterium tuberculosis (Véase, Groenen et al., Mol. Microbiol, 10:1057-1065 [1993]; Hoe et al., Emerg. Infect. Dis„ 5:254-263 [1999]; Masepohl et al., Biochim. Biophys. Acta 1307:26-30 [1996]; y Mojica et al., Mol. Microbiol, 17:85-93 [1995]). Los loci CRISPR normalmente difieren de otras SSR por la estructura de las repeticiones, que se han denominado repeticiones cortas y regularmente espaciadas (SRSR) (Janssen et al., OMICS J. Integ. Biol, 6:23-33 [2002]; y Mojica et al., Mol. Microbiol, 36:244-246 [2000]). En general, las repeticiones son elementos cortos que se producen en grupos que están espaciados regularmente por secuencias intermedias únicas con una longitud sustancialmente constante (Mojica et al., [2000], citado anteriormente). Aunque las secuencias de repetición están altamente conservadas entre las cepas, el número de repeticiones intercaladas y las secuencias de las regiones espaciadoras generalmente difieren de una cepa a otra (van Embden et al., J, Bacteriol, 182:2393-2401 [2000]). Los loci CRISPR se han identificado en más de 40 procariotas (véase, por ejemplo, Jansen et al., Mol. Microbiol, 43:1565-1575 [2002]; y Mojica et al., [2005]) que incluyen, pero sin limitación a Aeropyrum, Pyrobaculum, Sulfolobus, Archaeoglobus, Halocarcula, Methanobacterium, Methanococcus, Methanosarcina, Methanopyrus, Pyrococcus, Picrophilus, Thermoplasma, Corynebacterium, Mycobacterium, Streptomyces, Aquifex, Porphyromonas, Chlorobium, Thermus, Bacillus, Listeria, Staphylococcus, Clostridium, Thermoanaerobacter, Mycoplasma, Fusobacterium, Azarcus, Chromobacterium, Neisseria, Nitrosomonas, Desulfovibrio, Geobacter, Myxococcus, Campylobacter, Wolinella, Acinetobacter, Erwinia, Escherichia, Legionella,

Methylococcus, Pasteurella, Photobacterium, Salmonella, Xanthomonas, Yersinia, Treponema y Thermotoga.

En general, "Sistema CRISPR" se refiere colectivamente a transcritos y otros elementos involucrados en la expresión de o en dirigir la actividad de los genes asociados a CRISPR ("Cas"), incluidas las secuencias que codifican un gen Cas, una secuencia tracr (CRISPR transactivadora) (por ejemplo, ARNtracr o un ARNtracr parcial activo), una secuencia de apareamiento con tarc (que abarca una "repetición directa" y una repetición directa parcial procesada por ARNtracr en el contexto de un sistema CRISPR endógeno), una secuencia guía (también denominada "espaciador" en el contexto de un sistema CRISPR endógeno), u otras secuencias y transcritos de un locus CRISPR. Las expresiones secuencia guía y ARN guía se usan indistintamente. En algunas realizaciones, uno o más elementos de un sistema CRISPR derivan de un sistema CRISPR de tipo I, tipo II o tipo III. En algunas realizaciones, uno o más elementos de un sistema CRISPR derivan de un organismo particular que comprende un sistema CRISPR endógeno, tal como Streptococcus pyogenes. En general, un sistema CRISPR se caracteriza por elementos que promueven la formación de un complejo CRISPR en el sitio de una secuencia diana (también conocido como protoespaciador en el contexto de un sistema CRISPR endógeno). En el contexto de la formación de un complejo CRISPR, "secuencia diana" se refiere a una secuencia para la cual una secuencia guía se diseña para tener complementariedad, donde la hibridación entre una secuencia diana y una secuencia guía promueve la formación de un complejo CRISPR. Una secuencia diana puede comprender cualquier polinucleótido, tales como polinucleótidos de ADN o ARN. En algunas realizaciones, una secuencia diana se encuentra en el núcleo o citoplasma de una célula.

En las realizaciones preferidas, el sistema CRISPR es un sistema CRISPR tipo II y la enzima Cas es Cas9, que cataliza la escisión del ADN. La acción enzimática de Cas9 derivada de Streptococcus pyogenes o cualquier Cas9 estrechamente relacionada genera roturas bicatenarias en las secuencias del sitio diana que hibridan a 20 nucleótidos de la secuencia guía y tienen una secuencia de motivo adyacente al protoespaciador (PAM) NGG después de los 20 nucleótidos de la secuencia diana. La actividad de CRISPR a través de Cas9 para el reconocimiento y escisión del ADN específico del sitio se define por la secuencia guía, la secuencia tracr que hibrida en parte con la secuencia guía y la secuencia PAM. Más aspectos del sistema CRISPR se describen en Karginov y Hannon, The CRISPR system: small RNA-guided defense in bacteria and archae, Mole Cell 2010, January 15; 37(1): 7.

El locus de CRISPR tipo II de Streptococcus pyogenes SF370, contiene un grupo de cuatro genes Cas9, Cas1, Cas2 y Csn1, así como dos elementos de ARN no codificantes, ARNtracr y una matriz característica de secuencias repetitivas (repeticiones directas) interespaciadas por tramos cortos de secuencias no repetitivas (espaciadores, aproximadamente 30 pb cada uno). En este sistema, la rotura bicatenaria (DSB) de ADN dirigida se genera en cuatro pasos secuenciales. En primer lugar, dos ARN no codificantes, la matriz pre-ARNcr y el ARNtracr, se transcriben desde el locus CRISPR. En segundo lugar, se hibrida ARNtracr a las repeticiones directas de pre-ARNcr, que se procesa, a continuación, en ARNcr maduros que contienen secuencias espaciadoras individuales. En tercer lugar, el complejo de ARNcr:ARNtracr maduro dirige a Cas9 a la diana de ADN que consiste en el protoespaciador y al PAM correspondiente a través de la formación de heterodúplex entre la región de espaciador del ARNcr y el ADN protoespaciador. Por último, Cas9 media la escisión del ADN diana cadena arriba de PAM para crear un DSB dentro del protoespaciador. Varios aspectos del sistema CRISPR pueden mejorarse aún más para aumentar la eficacia y la versatilidad del direccionamiento de CRISPR. La actividad óptima de Cas9 puede depender de la disponibilidad de Mg2+ libre a niveles más altos que los presentes en el núcleo de mamíferos (véase, por ejemplo, Jinek et al., 2012, Science, 337:816) y la preferencia por un motivo NGG inmediatamente después del protoespaciador restringe la capacidad de dirigir en promedio cada 12 pb en el genoma humano.

Normalmente, en el contexto de un sistema CRISPR endógeno, la formación de un complejo CRISPR (que comprende una secuencia guía hibridada con una secuencia diana y complejada con una o más proteínas Cas) da como resultado la escisión de una o ambas cadenas en o cerca de (por ejemplo, en 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 50 o más pares de bases de) la secuencia diana. Sin pretender quedar ligados a teoría alguna, la secuencia tracr, que puede comprender o consistir en la totalidad o una porción de una secuencia tracr de tipo silvestre (por ejemplo, aproximadamente o aproximadamente más de 20, 26, 32, 45, 48, 54, 63, 67, 85 o más nucleótidos de una secuencia tracr de tipo silvestre), también puede formar parte de un complejo CRISPR, tal como por hibridación a lo largo de al menos una porción de la secuencia tracr con toda o una porción de una secuencia de apareamiento con tarc que está operativamente unida a la secuencia guía. En algunas realizaciones, uno o más vectores que dirigen la expresión de uno o más elementos de un sistema CRISPR se introducen en una célula hospedadora, de manera que la expresión de los elementos del sistema CRISPR dirige la formación de un complejo CRISPR en uno o más sitios diana. Por ejemplo, una enzima Cas, una secuencia guía unida a una secuencia de apareamiento con tarc y una secuencia tracr podrían unirse operativamente a elementos reguladores separados en vectores separados. Como alternativa, dos o más de los elementos expresados desde el mismo o diferentes elementos reguladores, pueden combinarse en un solo vector, con uno o más vectores adicionales que proporcionan cualquier componente del sistema CRISPR no incluido en el primer vector, elementos del sistema CRISPR que se combinan en un solo vector pueden estar dispuestos en cualquier orientación adecuada, como un elemento ubicado 5 'con respecto a ("cadena arriba" de) o 3' con respecto a ("cadena abajo" de) un segundo elemento. La secuencia de codificación de un elemento se puede ubicar en la misma cadena o en la cadena opuesta de la secuencia codificante de un segundo elemento y se puede orientar en la misma dirección u opuesta. En algunas realizaciones, un solo promotor dirige la expresión de un transcrito que codifica una enzima CRISPR y una o más de las secuencias guía, la secuencia de apareamiento con tarc (opcionalmente unida a la secuencia guía) y una secuencia tracr incrustada dentro de una o más secuencias intrón (por ejemplo, cada una en

un intrón diferente, dos o más en al menos un intrón, o todas en un solo intrón). En algunas realizaciones, la enzima CRISPR, la secuencia guía, secuencia de apareamiento con tarc y la secuencia tracr se unen operativamente a y se expresan desde el mismo promotor.

En algunas realizaciones, un vector comprende uno o más sitios de inserción, como una secuencia de reconocimiento de endonucleasas de restricción (también conocida como un "sitio de clonación"). En algunas realizaciones, uno o más sitios de inserción (por ejemplo, aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más sitios de inserción) están ubicados cadena arriba y/o cadena abajo de uno o más elementos de secuencia de uno o más vectores. En algunas realizaciones, un vector comprende un sitio de inserción cadena arriba de una secuencia de apareamiento con tarc y, opcionalmente, cadena abajo de un elemento regulador operativamente unido a la secuencia de apareamiento con tarc, de modo que después de la inserción de una secuencia guía en el sitio de inserción y tras la expresión, la secuencia guía dirige la unión específica de secuencia de un complejo CRISPR a una secuencia diana en una célula eucariota. En algunas realizaciones, un vector comprende dos o más sitios de inserción, cada sitio de inserción se ubica entre dos secuencias de apareamiento con tarc para permitir la inserción de una secuencia guía en cada sitio. En tal disposición, las dos o más secuencias guía pueden comprender dos o más copias de una única secuencia guía, dos o más secuencias guía diferentes, o combinaciones de éstas. Cuando se usan múltiples secuencias guía diferentes, se puede usar una construcción de expresión única para dirigir la actividad CRISPR a múltiples secuencias diana correspondientes diferentes dentro de una célula. Por ejemplo, un solo vector puede comprender aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, o más secuencias guía. En algunas realizaciones, pueden proporcionarse aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, o más, de dichos vectores que contienen secuencias guía y suministrarse opcionalmente a una célula.

En algunas realizaciones, un vector comprende un elemento regulador unido operativamente a una secuencia codificante de enzimas que codifica una enzima CRISPR, tal como una proteína Cas. Ejemplos no limitantes de proteínas Cas incluyen Cas1, Cas1B, Cas2, Cas3, Cas4, Cas5, Cas6, Cas7, Cas8, Cas9 (también conocida como Csn1 y Csx12), Cas10, Csy1, Csy2, Csy3, Cse1, Cse2, Csc1, Csc2, Csa5, Csn2, Csm2, Csm3, Csm4, Csm5, Csm6, Cmr1, Cmr3, Cmr4, Cmr5, Cmr6, Csb1, Csb2, Csb3, Csx17, Csx14, Csx10, Csx16, CsaX, Csx3, Csx1, Csx15, Csf1, Csf2, Csf3, Csf4, homólogos de los mismos, o versiones modificadas de los mismos. En algunas realizaciones, la enzima CRISPR sin modificar tiene actividad de escisión del ADN, tal como Cas9. En algunas realizaciones, la enzima CRISPR dirige la escisión de una o ambas cadenas en la ubicación de una secuencia diana, tal como dentro de la secuencia diana y/o dentro del complemento de la secuencia diana. En algunas realizaciones, la enzima CRISPR dirige la escisión de una o ambas cadenas dentro de aproximadamente 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, 100, 200, 500 o más pares de bases desde el primer o último nucleótido de una secuencia diana. En algunas realizaciones, un vector codifica una enzima CRISPR que está mutada con respecto a una enzima de tipo silvestre correspondiente, de manera que la enzima CRISPR mutada carece de la capacidad de escindir una o ambas cadenas de un polinucleótido diana que contiene una secuencia diana. Por ejemplo, una sustitución aspartato-a-alanina (D10A) en el dominio catalítico RuvC 1 de Cas9 de S. pyogenes convierte Cas9 de una nucleasa que escinde ambas cadenas a una nickasa (escinde una sola cadena). Otros ejemplos de mutaciones que hacen de Cas9 una nickasa incluyen, sin limitación, H840A, N854A y N863A, Como ejemplo adicional, dos o más dominios catalíticos de Cas9 (RuvC I, RuvC II y RuvC III o el dominio HNH) pueden mutarse para producir una Cas9 mutada que carece sustancialmente de toda la actividad de escisión del ADN. En algunas realizaciones, una mutación D10A se combina con uno o más de las mutaciones H840A, N854A o N863A para producir una enzima Cas9 que carece sustancialmente de toda la actividad de escisión del ADN. En algunas realizaciones, se considera que una enzima CRISPR carece sustancialmente de toda la actividad de escisión del ADN cuando la actividad de escisión del ADN de la enzima mutada es inferior a aproximadamente el 25%, del 10 %, del 5 %, del 1 %, del 0,1 %, el 0,01% o menos con respecto a su forma no mutada. Una sustitución aspartato-a-alanina (D10A) en el dominio catalítico RuvC I de SpCas9 convierte la nucleasa en una nickasa (véase, por ejemplo, Sapranauskas et al., 2011, Nucleic Acids Research, 39: 9275; Gasiunas et al., 2012, Proc. Natl. Acad. Sci. EE.UU., 109:E2579), de modo que el ADN genómico mellado se somete a la reparación dirigida por homología de alta fidelidad (HDR). En algunas realizaciones, una secuencia codificante de enzimas que codifica una enzima CRISPR es un codón optimizado para la expresión en células particulares, tales como células eucariotas. Las células eucariotas pueden ser de o derivadas de un organismo particular, tal como un mamífero, incluyendo, pero sin limitación de ser humano, ratón, rata, conejo, perro o primate no humano. En general, la optimización de codones se refiere a un proceso de modificación de una secuencia de ácido nucleico para una expresión mejorada en las células hospedadoras de interés reemplazando al menos un codón (por ejemplo, aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 10, 15, 20, 25, 50 o más codones) de la secuencia nativa con los codones que están con más frecuencia o que se usan con más frecuencia en los genes de esa célula hospedadora mientras se mantiene la secuencia de aminoácidos nativa. Varias especies muestran un sesgo particular para ciertos codones de un aminoácido particular. El sesgo de codón (las diferencias en el uso del codón entre organismos) a menudo se correlaciona con la eficacia de la traducción del ARN mensajero (ARNm), que a su vez se cree que depende de, entre otras cosas, las propiedades de los codones que se traducen y la disponibilidad de determinadas moléculas de ARN de transferencia (ARNt). El predominio de los ARNt seleccionados en una célula es generalmente un reflejo de los codones utilizados con mayor frecuencia en la síntesis de péptidos. Por consiguiente, los genes se pueden adaptar para una expresión génica óptima en un organismo dado en función de la optimización de codones. Las tablas de uso de codones están disponibles, véase Nakamura, Y., et al. “Codon usage tabulated from the international DNA sequence databases: status for the year 2000”Nucl. Acids Res. 28:292 (2000). También están disponibles los algoritmos informáticos para el codón que optimiza una secuencia particular para la expresión en una célula

hospedadora particular, tales como Gene Forge (Aptagen; Jacobus, PA), En algunas realizaciones, uno o más codones (por ejemplo, 1, 2, 3, 4, 5, 10, 15, 20, 25, 50 o más, o todos los codones) en una secuencia que codifica una enzima CRISPR corresponden al codón más frecuentemente usado para un determinado aminoácido.

En algunas realizaciones, un vector codifica una enzima CRISPR que comprende una o más secuencias de localización nuclear (NLS), tal como aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más NLS. En algunas realizaciones, la enzima CRISPR comprende aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más NLS en o cerca del extremo amino, aproximadamente o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más NLS en o cerca del extremo carboxi, o una combinación de estas (por ejemplo, una o más NLS en el extremo amino y una o más NLS en el extremo carboxi). Cuando hay más de una NLS, cada una puede seleccionarse independientemente de las otras, de modo que una sola NLS puede estar presente en más de una copia y/o en combinación con una o más NLS presentes en una o más copias. En una realización preferida, la enzima CRISPR comprende a lo sumo 6 NLS. En algunas realizaciones, una NLS se considera cerca del extremo N o C cuando el aminoácido más cercano de la NLS está dentro de aproximadamente 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50 o más amino ácidos a lo largo de la cadena polipeptídica del extremo N o C. Ejemplos no limitantes de NLS incluyen una secuencia NLS derivada de: la NLS del antígeno T grande del virus SV40, que tiene la secuencia de aminoácidos PKKKRKV; la NLS de nucleoplasmina (por ejemplo, la NLS de nucleoplasmina bipartita con la secuencia KRPAATKKAGQAKKKK); la NLS de c-myc que tiene la secuencia de aminoácidos PAAKRVKLD o RQRRNELKRSP; la NLS de hRNPA1 M9 tiene la secuencia NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY; la secuencia RMRIZFKNKGKDTAELRRRVEVSVELRKAKKDEQILKRRNV del dominio IBB de importina alfa; las secuencias VSRKRPRP y PPKKARED de la proteína T de mioma; la secuencia POPKKKPL de p53 humana; la secuencia SALIKKKKKMAP c-abl IV de ratón; las secuencias DRLRR y PKQKKRK del virus de la influenza NS1; la secuencia RKLKKKIKKL del antígeno delta del virus de la hepatitis; la secuencia REKKKFLKRR de la proteína Mx1 de ratón; la secuencia KRKGDEVDGVDEVAKKKSKK de la poli (ADP-ribosa) polimerasa humana; y la secuencia RKCLQAGMNLEARKTKK de los receptores de hormonas esteroideas (humanos) glucocorticoides.

En general, uno o más NLS tienen la fuerza suficiente para impulsar la acumulación de la enzima CRISPR en una cantidad detectable en el núcleo de una célula eucariota. En general, la fuerza de la actividad de localización nuclear puede derivar del número de secuencias de localización nuclear (NLS) en la enzima CRISPR, la(s) NLS particular(es) usada(s), o una combinación de estos factores. La detección de la acumulación en el núcleo se puede realizar mediante cualquier técnica adecuada. Por ejemplo, un marcador detectable puede fusionarse con la enzima CRISPR, de modo que pueda visualizarse la ubicación dentro de una célula, tal como en combinación con un medio para detectar la ubicación del núcleo (por ejemplo, una tinción específica para el núcleo como DAPI). Los núcleos celulares también pueden aislarse de las células, cuyos contenidos pueden analizarse a continuación mediante cualquier proceso adecuado para detectar proteínas, como inmunohistoquímica, transferencia Western o ensayo de actividad enzimática. La acumulación en el núcleo también puede determinarse indirectamente, como por ejemplo mediante un ensayo para determinar el efecto de la formación del complejo CRISPR (por ejemplo, un ensayo para determinar la escisión o mutación del ADN en la secuencia diana, o un ensayo para determinar la actividad de expresión génica alterada afectada por la formación del complejo CRISPR y/o la actividad de la enzima CRISPR), en comparación con un control no expuesto a la enzima o complejo CRISPR, o expuesto a una enzima CRISPR que carece de una o más NLS.

En general, una secuencia guía es cualquier secuencia de polinucleótidos que tenga suficiente complementariedad con una secuencia de polinucleótidos diana para hibridar con la secuencia diana y dirigir la unión específica de la secuencia de un complejo CRISPR a la secuencia diana. A lo largo de esta solicitud, la secuencia guía se puede referir indistintamente como una guía o un espaciador. En algunas realizaciones, el grado de complementariedad entre una secuencia guía y su secuencia diana correspondiente, cuando se alinea de manera óptima utilizando un algoritmo de alineamiento adecuado, es aproximadamente o aproximadamente más del 50%, del 60 %, del 75 %, del 80 %, del 85 %, del 90 %, del 95 %, del 97,5 %, el 99 % o más. El alineamiento óptimo se puede determinar con el uso de cualquier algoritmo adecuado para alinear secuencias, cuyo ejemplo no limitante incluye el algoritmo Smith-Waterman, el algoritmo Needleman-Wunsch, algoritmos basados en la Transformación de Burrows-Wheeler (por ejemplo, el Alineador Wheeler Burrows), ClustalW, Clustal X, BLAT, Novoalign (Novocraft Technologies; disponible en www.novocraft.com), ELAND (Illumina, San Diego, CA), SOAP (disponible en soap.genomics.org.cn), y Maq (disponible en maq.sourceforge.net). En algunas realizaciones, la secuencia diana candidata tiene una longitud de 5, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 75, o más nucleótidos de longitud. En algunas realizaciones, una secuencia guía es menor de aproximadamente 75, 50, 45, 40, 35, 30, 25, 20, 15, 12 o menos nucleótidos de longitud. La capacidad de una secuencia guía para dirigir la unión específica de la secuencia de un complejo CRISPR a una secuencia diana se puede evaluar mediante cualquier ensayo adecuado. Por ejemplo, los componentes de un sistema CRISPR suficientes para formar un complejo CRISPR, incluyendo la secuencia guía a probar, pueden proporcionarse a una célula hospedadora que tenga la secuencia diana correspondiente, tal como por transfección con vectores que codifican los componentes de la secuencia CRISPR, seguida de una evaluación de la escisión preferencial dentro de la secuencia diana, tal como mediante el ensayo SURVEYOR como se describe en el presente documento. De manera similar, la escisión de una secuencia de polinucleótidos diana puede evaluarse en un tubo de ensayo proporcionando la secuencia diana, los componentes de un complejo CRISPR, incluida la secuencia guía a probar y una secuencia guía de control diferente de la secuencia guía de prueba, y comparando la unión o tasa de escisión en la secuencia diana entre las reacciones de la secuencia

guía de prueba y de control. Son posibles otros ensayos y se les ocurrirán a los expertos en la materia.

Se puede seleccionar una secuencia guía para dirigir a cualquier secuencia diana. En algunas realizaciones, la secuencia diana es una secuencia dentro de un genoma de una célula. Las secuencias diana a modo de ejemplo incluyen aquellas que son únicas en el genoma diana. Por ejemplo, para la Cas9 de S. pyogenes, una secuencia diana única en un genoma puede incluir un sitio diana de Cas9 de la forma MMMMMMMMNNNNNNNNNNNNXGG dondeNNNNNNNNNNNNXGG (N es A, G, T, o C; y X puede ser cualquiera) tiene una única aparición en el genoma. Una secuencia diana única en un genoma puede incluir un sitio diana de Cas9 de S. pyogenes de la forma MMMMMMMMMNNNNNNNNNNNNXGG donde NNNNNNNNNNNXGG (N es A, G, T, o C; y X puede ser cualquiera) tiene una única aparición en el genoma. Para el CRISPR1 Cas9 S. thermophilus, una secuencia diana única en un genoma puede incluir un sitio diana de Cas9 de la forma MMMMMMMMNNNNNNNNNNNNXXAGAAW donde NNNNNNNNNNNNXXAGAAW (N es A, G, T, o C; X puede ser cualquiera; yW es A o T) tiene una única aparición en el genoma. Una secuencia diana única en un genoma puede incluir un sitio diana de CRISPR1 Cas9 de S. thermophilus de la forma MMMMMMMMMNNNNNNNNNNNNXXAGAAW donde NNNNNNNNNNNXXAGAAW (N es A, G, T, o C; X puede ser cualquiera; y W es A o T) tiene una única aparición en el genoma. Para la Cas9 de S. pyogenes, una secuencia diana única en un genoma puede incluir un sitio diana de Cas9 de la forma MMMMMMMMNNNNNNNNNNNNXGGXG donde NNNNNNNNNNNNXGGXG (N es A, G, T, o C; y X puede ser cualquiera) tiene una única aparición en el genoma. Una secuencia diana única en un genoma puede incluir un sitio diana de Cas9 de S. pyogenes de la forma MMMMMMMMMNNNNNNNNNNNXGGXG donde NNNNNNNNNNNXGGXG (N es A, G, T, o C; y X puede ser cualquiera) tiene una única aparición en el genoma. En cada una de estas secuencias, "M" puede ser A, G, T o C, y no es necesario considerarlas al identificar una secuencia como única.

En algunas realizaciones, se selecciona una secuencia guía para reducir el grado de estructura secundaria dentro de la secuencia guía. En algunas realizaciones, aproximadamente o aproximadamente menos del 75 %, del 50 %, del 40 %, del 30 %, del 25 %, del 20 %, del 15 %, del 10 %, del 5 %, del 1% o menos de los nucleótidos de la secuencia guía participan en el emparejamiento de bases auto-complementarias cuando se pliegan de manera óptima. El plegamiento óptimo puede determinarse mediante cualquier algoritmo de plegamiento de polinucleótidos adecuado. Algunos programas se basan en el cálculo de la energía libre mínima de Gibbs. Un ejemplo de uno de estos algoritmos es mFold, como se describe por Zuker y Stiegler (Nucleic Acids Res. 9 (1981), 133-148). Otro ejemplo de algoritmo de plegamiento es el servidor web en línea RNAfold, desarrollado en el `Instituto de Química Teórica de la Universidad de Viena´, que utiliza el algoritmo de predicción de la estructura del centroide (véase, por ejemplo, A.R. Gruber et al., 2008, Cell 106(1): 23-24; y PA Carr y GM Church, 2009, Nature Biotechnology 27(12): 1151 -62).

En general, una secuencia de apareamiento con tarc incluye cualquier secuencia que, tiene suficiente complementariedad con una secuencia tracr para promover una o más de: (1) la escisión de una secuencia guía flanqueada por secuencias de apareamiento con tarc en una célula que contiene la secuencia tracr correspondiente; y (2) la formación de un complejo CRISPR en una secuencia diana, en donde el complejo CRISPR comprende la secuencia de apareamiento con tarc hibridada a la secuencia tracr. En general, el grado de complementariedad es con referencia al alineamiento óptimo de la secuencia de apareamiento con tarc y la secuencia tracr, a lo largo de la longitud de la más corta de las dos secuencias. El alineamiento óptimo puede determinarse mediante cualquier algoritmo de alineamiento adecuado, y puede explicar además las estructuras secundarias, tales como la auto complementariedad dentro de la secuencia tracr o de la secuencia de apareamiento con tarc. En algunas realizaciones, el grado de complementariedad entre la secuencia tracr y la secuencia de apareamiento con tarc a lo largo de la longitud de la más corta de las dos cuando se alinean de manera óptima es aproximadamente o aproximadamente más del 25%, del 30 %, del 40 %, del 50 %, del 60 %, del 70 %, del 80 %, del 90 %, del 95 %, del 97,5 %, del 99 % o mayor. En algunas realizaciones, la secuencia tracr es aproximadamente o aproximadamente más de 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 25, 30, 40, 50, o más nucleótidos de longitud. En algunas realizaciones, la secuencia tracr y la secuencia de apareamiento con tarc están contenidas dentro de un solo transcrito, de modo que la hibridación entre las dos produce un transcito que tiene una estructura secundaria, tal como una horquilla. En una realización de la divulgación, el transcrito o la secuencia de polinucleótidos transcrita tiene al menos dos o más horquillas. En las realizaciones preferidas, el transcrito tiene dos, tres, cuatro o cinco horquillas. En una realización adicional de la divulgación, el transcrito tiene a lo sumo cinco horquillas. En una estructura en forma de horquilla, la parte de la secuencia 5 ‘ del "N" final y cadena arriba del bucle corresponde a la secuencia de apareamiento con tarc, y la parte de la secuencia 3' del bucle corresponde a la secuencia tracr. En la parte inferior de la Figura 15B se proporciona una ilustración de ejemplo de dicha estructura de horquilla. Otros ejemplos no limitantes de polinucleótidos individuales que comprenden una secuencia guía, una secuencia de apareamiento con tarc y una secuencia tracr son los siguientes (enumerados de 5 'a 3'), donde "N" representa una base de una secuencia guía, el primer bloque de las letras minúsculas representan la secuencia de apareamiento con tarc, y el segundo bloque de letras minúsculas representa la secuencia tracr, y la secuencia poli-T final representa el terminador de la transcripción: (1) NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataaggctt catgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT; (2) NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatca acaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT; (3) NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataag gcttcatgccgaaatca acaccctgtcattttatggcagggtgtTITTTT; (4)

NNNNNNNNNNNNNNNNNNNNgttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaacttgaaaa agtggcaccgagtcggtgcTTTTTT; (5)
NNNNNNNNNNNNNNNNNNNNgttttagagctaGAAATAGcaagttaaaataaggctagtccgttatcaacttgaa aaagtgTTTTTTT; y (6)
NNNNNNNNNNNNNNNNNNNNgttttagagctagAAATAGcaagttaaaataaggctagtccgttatcaTTTTT TTT. En algunas realizaciones, las secuencias (1) a (3) se utilizan en combinación con Cas9 de CRISPR1 de S. thermophilus. En algunas realizaciones, las secuencias (4) a (6) se usan en combinación con Cas9 de S. pyogenes. En algunas realizaciones, la secuencia tracr es un transcrito separado de un transcrito, que comprende la secuencia de apareamiento con tracr.

En algunas realizaciones, se proporciona también una plantilla de recombinación. Una plantilla de recombinación puede ser un componente de otro vector como se describe en el presente documento, contenido en un vector separado, o proporcionado como un polinucleótido separado. En algunas realizaciones, una plantilla de recombinación se diseña para servir como plantilla en la recombinación homóloga, tal como dentro o cerca de una secuencia diana mellada o escindida por una enzima CRISPR como una parte, de un complejo CRISPR. Un polinucleótido plantilla puede tener cualquier longitud adecuada, tal como aproximadamente o aproximadamente más de 10, 15, 20, 25, 50, 75, 100, 150, 200, 500, 1000 o más nucleótidos de longitud. En algunas realizaciones, el polinucleótido plantilla es complementario a una porción de un polinucleótido que comprende la secuencia diana. Cuando se alinean de forma óptima, un polinucleótido plantilla podría superponerse con uno o más nucleótidos de unas secuencias diana (por ejemplo, aproximadamente o aproximadamente más de 1, 5, 10, 15, 20 o más nucleótidos), en algunas realizaciones, cuando una secuencia plantilla y un polinucleótido que comprende una secuencia diana están alineados de manera óptima, el nucleótido más cercano del polinucleótido plantilla está dentro de aproximadamente 1, 5, 10, 15, 20, 25, 50, 75, 100, 200, 300, 400, 500, 1000, 5000, 10000 o más nucleótidos de la secuencia diana.

En algunas realizaciones, la enzima CRISPR es parte de una proteína de fusión que comprende uno o más dominios de proteínas heterólogas (por ejemplo, aproximadamente, o aproximadamente más de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más dominios además de la enzima CRISPR). Una proteína de fusión enzimática CRISPR puede comprender cualquier secuencia de proteínas adicional y, opcionalmente, una secuencia de enlace entre dos dominios cualesquiera. Ejemplos de dominios de proteínas que pueden fusionarse con una enzima CRISPR incluyen, sin limitación, etiquetas epitópicas, secuencias de genes indicadores y dominios de proteínas que tienen una o más de las siguientes actividades: actividad metilasa, actividad demetilasa, actividad de activación de la transcripción, actividad de represión de la transcripción, actividad de factor de liberación de la transcripción, actividad de modificación de histonas, Actividad de escisión del ARN y actividad de unión al ácido nucleico. Ejemplos no limitantes de etiquetas de epítopo incluyen etiquetas de histidina (His), etiquetas V5, etiquetas FLAG, etiquetas de hemaglutinina de influenza (HA), Etiquetas Myc, Etiquetas VSV-G y etiquetas tiorredoxina (Trx). Ejemplos de genes indicadores incluyen, pero sin limitación, glutatión-S-transferasa (GST), peroxidasa de rábano picante (HRP), cloranfenicol acetiltransferasa (CAT) beta-galactosidasa, beta-glucuronidasa, luciferasa, proteína fluorescente verde (GFP), HcRed, DsRed, proteína fluorescente cian (CFP), proteína fluorescente amarilla (YFP) y proteínas autofluorescentes, incluida la proteína fluorescente azul (BFP). Una enzima CRISPR se puede fusionar con una secuencia de genes que codifica una proteína o un fragmento de una proteína que se une a moléculas de ADN o se une a otras moléculas celulares, incluidas, pero sin limitación, la proteína de unión a maltosa (MBP), etiqueta S, fusiones de dominio de unión al ADN Lex A (DBD), fusiones de dominio de unión a ADN GAL4 y fusiones de proteínas BP 16 del virus herpes simple (VHS). Los dominios adicionales que pueden formar parte de una proteína de fusión que comprende una enzima CRISPR se describen en el documento US20110059502. En algunas realizaciones, se utiliza una enzima CRISPR etiquetada para identificar la ubicación de una secuencia diana.

En algunas realizaciones, una enzima CRISPR puede formar un componente de un sistema inducible. La naturaleza inducible del sistema permitiría el control espacio temporal de la edición de genes o de la expresión de genes utilizando una forma de energía. La forma de energía puede incluir, pero sin limitación, radiación electromagnética, energía acústica, energía química y energía térmica. Ejemplos de sistema inducible incluyen promotores inducibles por tetraciclina (Tet-On o Tet-Off), sistemas de activación de transcripción de dos híbridos de molécula pequeña (FKBP, ABA, etc.), o sistemas inducibles por luz (Fitocromo, dominios LOV, o criptocoroma). En una realización, la enzima CRISPR puede ser una parte de un efector transcripcional inducible por la luz (LITE) para dirigir los cambios en la actividad transcripcional de una manera específica de la secuencia. Los componentes de una luz pueden incluir una enzima CRISPR, un heterodímero citocromo sensible a la luz (por ejemplo, de Arabidopsis thaliana) y un dominio de activación/represión transcripcional. Se proporcionan ejemplos adicionales de proteínas de unión a ADN inducibles y métodos para su uso en los documentos US 61/736465 y US 61/721.283.

En algunos aspectos, la divulgación comprende entregar uno o más polinucleótidos, tales como o uno o más vectores como se describe en el presente documento, uno o más transcritos de los mismos, y/o una o más proteínas transcritas de los mismos, a una célula hospedadora. En algunos aspectos, la divulgación comprende las células producidas por tales métodos y los animales comprendidos o producidos a partir de dichas células. En algunas realizaciones, se administra una enzima CRISPR en combinación con (y opcionalmente complejada con) una secuencia guía a una célula. Se pueden usar métodos convencionales de transferencia de genes basados en virus y en no virus para introducir ácidos nucleicos en células o tejidos diana de mamíferos. Tales métodos pueden usarse para administrar ácidos nucleicos que codifican componentes de un sistema CRISPR a células en cultivo, o en el organismo hospedador. Los sistemas de administración de vectores no virales incluyen plásmidos de ADN, ARN (por ejemplo, un

transcrito de un vector descrito en le presente documento), ácido nucleico desnudo y ácido nucleico complejado con un vehículo de administración, tal como un liposoma. Los sistemas de administración de vectores virales incluyen virus de ADN y ARN, que tienen genomas episómicos o integrados después de la administración a la célula. Para una revisión de los procedimientos de terapia génica, véase Anderson, Science 256:808-813 (1992); Nabel & Feigner, TIBTECH 11:211-217 (1993); Mitani & Caskey, TIBTECH 11:162-166 (1993); Dillon, TIBTECH 11:167-175 (1993); Miller, Nature 357:455-460 (1992); Van Brunt, Biotechnology 6(10):1149-1154 (1988); Vigne, Restorative Neurology and Neuroscience 8:35-36 (1995); Kremer & Perricaudet, British Medical Bulletin 51(1):31-44 (1995); Haddada et al., in Current Topics in Microbiology and Immunology Doerfler and Böhm (eds) (1995); y Yu et al., Gene Therapy 1:13-26 (1994).

En algunas realizaciones, una célula hospedadora contiene la secuencia diana, y la célula puede proceder de células tomadas de un sujeto, tal como una línea celular. Se conoce una amplia diversidad de líneas celulares para el cultivo de tejidos en la materia. Los ejemplos de líneas celulares incluyen, pero sin limitación, C8161, CCRF-CEM, MOLT, mIMCD-3, NHDF, HeLa-S3, Huh1, Huh4, Huh7, HUVEC, HASMC, HEKn, HEKa, MiaPaCell, Panc1, PC-3, TF1, CTLL-2, C1R, Rat6, CV1, RPTE, A10, T24, J82, A375, ARH-77, Calu1, SW480, SW620, SKOV3, SK-UT, CaCo2, P388D1, SEM-K2, WEHI-231, HB56, TIB55, Jurkat, J45.01, LRMB, Bcl-1, BC-3, IC21, DLD2, Raw264.7, NRK, NRK-52E, MRC5, MEF, Hep G2, HeLa B, HeLa T4, COS, COS-1, COS-6, COS-M6A, BS-C-1 epitelial de riñón de mono, BALB/3T3 de fibroblasto de embrión de ratón, 3T3 Swiss, 3T3-L1, 132-d5 de fibroblastos fetales humanos; 10.1 de fibroblastos de ratón, 293-T, 3T3, 721, 9L, A2780, A2780ADR, A2780cis, A172, A20, A253, A431, A-549, ALC, B16, B35, células BCP-1, BEAS-2B, bEnd.3, BHK-21, BR 293, BxPC3, C3H-10T1/2, C6/36, Cal-27, CHO, CHO-7, CHO-IR, CHO-K1, CHO-K2, CHO-T, CHO Dhfr -/-, COR-L23, COR-L23/CPR, COR-L23/5010, COR-L23/R23, COS-7, COV-434, CML T1, CMT, CT26, D17, DH82, DU145, DuCaP, EL4, EM2, EM3, EMT6/AR1, EMT6/AR10.0, FM3, H1299, H69, HB54, HB55, HCA2, HEK-293, HeLa, Hepalclc7, HL-60, HMEC, HT-29, Jurkat, células JY, células K562, Ku812, KCL22, KG1, KYO1, LNCap, Ma-Mel 1-48, MC-38, MCF-7, MCF-10A, MDA-MB-231, MDA-MB-468, MDA-MB-435, MDCK II, MDCK II, MOR/0.2R, MONO-MAC 6, MTD-1A, MyEnd, NCI-H69/CPR, NCI-H69/LX10, NCI-H69/LX20, NCIH69/ LX4, NIH-3T3, NALM-1, NW-145, líneas celulares OPCN / OPCT, Peer, PNT-1A / PNT 2, RenCa, RIN-5F, RMA/RMAS, células Saos-2, Sf-9, SkBr3, T2, T-47D, T84, línea celular THP1, U373, U87, U937, VCaP, células Vero, WM39, WT-49, X63, YAC-1, YAR, y variantes transgénicas de las mismas. Las líneas celulares están disponibles en varias fuentes conocidas por los expertos en la materia (véase, por ejemplo, La Colección Americana de Cultivos Tipo (ATCC) (Manassus, Va.)). En algunas realizaciones, una célula transfectada con uno o más vectores descritos en el presente documento se usa para establecer una nueva línea celular que comprende una o más secuencias procedentes de vectores. En algunas realizaciones, una célula transfectada transitoriamente con los componentes de un sistema CRISPR como se describe en el presente documento (tal como mediante transfección transitoria de uno o más vectores, o transfección con ARN), y modificada a través de la actividad de un complejo CRISPR, se usa para establecer una nueva línea celular que comprende células que contienen la modificación pero que carecen de cualquier otra secuencia exógena. En algunas realizaciones, las células transfectadas de forma transitoria o no transitoria con uno o más vectores descritos en el presente documento, o las líneas celulares procedentes de tales células se usan para evaluar uno o más compuestos de prueba. Las secuencias diana pueden estar en dichas células.

Con los avances recientes en la genómica de cultivos, la capacidad de usar los sistemas CRISPR-Cas9 para realizar la manipulación y edición de genes eficaz y rentable permitirá la rápida selección y comparación de manipulaciones genéticas únicas y multiplexadas para transformar dichos genomas para mejorar la producción y mejorar las características. A este respecto se hace referencia a patentes y publicaciones de Estados Unidos: Patente de Estados Unidos Nº 6.603.061 - Agrobacterium-Mediated Plant Transformation Method; Patente de Estados Unidos N.º 7.868.149 - Plant Genome Sequences and Uses Thereof y US 2009/0100536 -Transgenic Plants with Enhanced Agronomic Traits- En la práctica de la invención, se hace referencia al contenido y la divulgación de Morrell et al. “Crop genomics:advances and applications” Nat Rev Genet. 2011 Dec 29;13(2):85-96.

En una realización ventajosa de la invención, el sistema CRISPR/Cas9 se utiliza para diseñar microalgas. Por lo tanto, los polinucleótidos diana en la invención pueden ser plantas, algas, procariotas o eucariotas.

Los sistemas CRISPR pueden ser útiles para crear un animal o célula que puede usarse como modelo de enfermedad. Por lo tanto, la identificación de secuencias diana para los sistemas CRISPR pueden ser útiles para crear un animal o célula que puede usarse como modelo de enfermedad. Como se usa en el presente documento, "enfermedad" se refiere a una enfermedad, trastorno o indicación en un sujeto. Por ejemplo, un método de la invención puede usarse para crear un animal o célula que comprende una modificación en una o más secuencias de ácido nucleico asociadas con una enfermedad, o un animal o célula en el que se altera la expresión de una o más secuencias de ácido nucleico asociadas con una enfermedad. Dicha secuencia de ácido nucleico puede codificar una secuencia de proteína asociada a enfermedad o puede ser una secuencia de control asociada a enfermedad.

En algunos métodos, el modelo de enfermedad se puede usar para estudiar los efectos de las mutaciones en el animal o la célula y el desarrollo y/o la progresión de la enfermedad utilizando medidas comúnmente utilizadas en el estudio de la enfermedad. Como alternativa, dicho modelo de enfermedad es útil para estudiar el efecto de un compuesto farmacéuticamente activo sobre la enfermedad.

En algunos métodos, el modelo de enfermedad se puede usar para evaluar la eficacia de una estrategia potencial de

terapia génica. Es decir, un gen o polinucleótido asociado a la enfermedad puede modificarse de tal manera que el desarrollo y/o la progresión de la enfermedad se inhiban o reduzcan. En particular, el método comprende modificar un gen o polinucleótido asociado con la enfermedad de manera que se produzca una proteína alterada y, como resultado, el animal o la célula tenga una respuesta alterada. Por consiguiente, en algunos métodos, un animal modificado genéticamente puede compararse con un animal predispuesto al desarrollo de la enfermedad de tal manera que se pueda evaluar el efecto del evento de terapia génica.

Los sistemas CRISPR se pueden usar para desarrollar un agente biológicamente activo que modula un evento de señalización celular asociado con un gen de la enfermedad; y por lo tanto, la identificación de secuencias diana se puede utilizar así.

Los sistemas CRISPR se pueden usar para desarrollar un modelo celular o el modelo animal se puede construir en combinación con el método de la invención para detectar un cambio de función celular; y por lo tanto, la identificación de secuencias diana se puede utilizar así. Dicho modelo puede usarse para estudiar los efectos de una secuencia genómica modificada mediante el complejo CRISPR en una función celular de interés. Por ejemplo, un modelo de función celular puede usarse para estudiar el efecto de una secuencia de genoma modificada en la señalización intracelular o señalización extracelular. Como alternativa, se puede usar un modelo de función celular para estudiar los efectos de una secuencia de genoma modificada en la percepción sensorial. En algunos de dichos modelos, se modifican una o más secuencias genómicas asociadas con una ruta bioquímica de señalización en el modelo.

Una expresión alterada de una o más secuencias genómicas asociadas con una ruta bioquímica de señalización puede determinarse mediante el análisis de una diferencia en los niveles de ARNm de los genes correspondientes entre la célula modelo de prueba y una célula de control, cuando se ponen en contacto con un agente candidato. Como alternativa, la expresión diferencial de las secuencias asociadas con una ruta bioquímica de señalización se determina mediante la detección de una diferencia en el nivel del polipéptido codificado o producto génico. Para ensayar una alteración inducida por un agente en el nivel de transcritos de ARNm o polinucleótidos correspondientes, se extrae primero el ácido nucleico contenido en una muestra de acuerdo con los métodos estándar en la materia. Por ejemplo, el ARNm puede aislarse utilizando varias enzimas líticas o soluciones químicas de acuerdo con los procedimientos establecidos en Sambrook et al. (1989), o se extraen mediante resinas de unión a ácidos nucleicos siguiendo las instrucciones de la compañía proveedora de CA proporcionadas por los fabricantes. Luego, se detecta el ARNm contenido en la muestra de ácido nucleico extraído mediante procedimientos de amplificación o ensayos de hibridación convencionales (por ejemplo, análisis de transferencia Northern) de acuerdo con métodos ampliamente conocidos en la materia o basados en los métodos ejemplificados en el presente documento.

Para los fines de la presente invención, amplificación significa cualquier método que emplee un cebador y una polimerasa capaces de replicar una secuencia diana con una fidelidad razonable. La amplificación se puede realizar mediante ADN polimerasas naturales o recombinantes, tales como TaqGold™, ADN polimerasa T7, fragmento Klenow de ADN polimerasa de E. coli y transcriptasa inversa, Un método de amplificación preferido es la PCR. En particular, el ARN aislado puede someterse a un ensayo de transcripción inversa que se acopla con una reacción en cadena de polimerasa cuantitativa (RT-PCR) para cuantificar el nivel de expresión de una secuencia asociada con una ruta bioquímica de señalización.

La detección del nivel de expresión génica se puede realizar en tiempo real en un ensayo de amplificación. En un aspecto, los productos amplificados pueden visualizarse directamente con agentes de unión a ADN fluorescentes que incluyen, pero no se limitan a, intercaladores de ADN y compuestos de unión al surco de ADN. Debido a que la cantidad de los intercaladores incorporados en las moléculas de ADN de doble cadena es, normalmente, proporcional a la cantidad de los productos de ADN amplificados, se puede determinar convenientemente la cantidad de los productos amplificados mediante la cuantificación de la fluorescencia del colorante intercalado utilizando sistemas ópticos convencionales. El colorante de unión al ADN adecuado para esta aplicación incluye SYBR verde, SYBR azul, DAPI, yoduro de propidio, Hoeste, SYBR oro, bromuro de etidio, acridinas, proflavina, naranja de acridina, acriflavina, fluorcoumanina, ellipticina, daunomicina, cloroquina, distamicina D, cromomicina, homidio, mitramicina, polipiridilos de rutenio, antramicina, y similares.

En otro aspecto, se pueden emplear otros marcadores fluorescentes tales como sondas específicas de secuencia en la reacción de amplificación para facilitar la detección y cuantificación de los productos amplificados. La amplificación cuantitativa basada en sondas se basa en la detección específica de secuencia de un producto amplificado deseado. Utiliza sondas fluorescentes, específicas de dianas (por ejemplo, sondas TaqMan®) que dan lugar a una mayor especificidad y sensibilidad. Los métodos para realizar la amplificación cuantitativa basada en sondas están bien establecidos en la materia y se enseñan en la Patente de Estados Unidos Nº 5.210.015.

En otro aspecto más, se pueden realizar ensayos de hibridación convencionales que usan sondas de hibridación que comparten homología de secuencia con secuencias asociadas con una ruta bioquímica de señalización. Normalmente, se permite que las sondas formen complejos estables con las secuencias asociadas con una ruta bioquímica de señalización contenida dentro de la muestra biológica procedente del sujeto de prueba en una reacción de hibridación. Un experto en la materia apreciará que cuando se usa antisentido como el ácido nucleico de la sonda, los polinucleótidos diana proporcionados en la muestra se eligen para que sean complementarios a las secuencias de los

ácidos nucleicos antisentido. Por el contrario, donde la sonda de nucleótidos es un ácido nucleico de sentido, el polinucleótido diana se selecciona para que sea complementario a las secuencias del ácido nucleico de sentido.

La hibridación se puede realizar en condiciones de varias restricciones. Las condiciones de hibridación adecuadas para la práctica de la presente invención son tales que la interacción de reconocimiento entre la sonda y las secuencias asociadas con una ruta bioquímica de señalización es tanto específica como suficientemente estable. Las condiciones que aumentan la rigurosidad de una reacción de hibridación son ampliamente conocidas y publicadas en la materia. Véase, por ejemplo, (Sambrook, et al., (1989); Nonradioactive In Situ Hybridization Application Manual, Boehringer Mannheim, segunda edición). El ensayo de hibridación se puede formar utilizando sondas inmovilizadas sobre cualquier soporte sólido, que incluye, pero sin limitación a, nitrocelulosa, vidrio, silicio y varias matrices genéticas. Un ensayo de hibridación preferido se lleva a cabo en chips genéticos de alta densidad como se describe en la Patente de Estados Unidos Nº 5.445.934.

Para una detección conveniente de los complejos de sonda-diana formados durante el ensayo de hibridación, las sondas de nucleótidos se conjugan a un marcador detectable. Los marcadores detectables adecuados para su uso en la presente invención incluyen cualquier composición detectable por medios fotoquímicos, bioquímicos, espectroscópicos, inmunoquímicos, eléctricos, ópticos o químicos. Se conocen en la materia una amplia variedad de marcadores detectables apropiados, que incluyen marcadores fluorescentes o quimioluminiscentes, marcadores de isótopos radiactivos, ligandos enzimáticos u otros. En las realizaciones preferidas, es probable que se desee emplear un marcador fluorescente o un marcador de enzima, tal como digoxigenina, β-galactosidasa, ureasa, fosfatasa alcalina o peroxidasa, complejo de avidina/biotina.

Los métodos de detección utilizados para detectar o cuantificar la intensidad de hibridación dependerán normalmente del marcador seleccionado anteriormente. Por ejemplo, los radiomarcadores pueden detectarse utilizando una película fotográfica o una placa de fósforo. Los marcadores fluorescentes se pueden detectar y cuantificar utilizando un fotodetector para detectar la luz emitida. Los marcadores enzimáticos se detectan normalmente proporcionando a la enzima un sustrato y midiendo el producto de reacción producido mediante la acción de la enzima en el sustrato; y finalmente los marcadores colorimétricos se detectan simplemente visualizando el marcador coloreado.

Un cambio inducido por un agente en la expresión de secuencias asociadas con una ruta bioquímica de señalización también se puede determinar mediante el examen de los productos génicos correspondientes. La determinación del nivel de proteína generalmente implica a) poner en contacto la proteína contenida en una muestra biológica con un agente que se une específicamente a una proteína asociada con una ruta bioquímica de señalización; y (b) identificar cualquier complejo agente:proteína así formado. En un aspecto de esta realización, el agente que se une específicamente a una proteína asociada con una ruta bioquímica de señalización es un anticuerpo, preferentemente un anticuerpo monoclonal. La reacción se realiza poniendo en contacto el agente con una muestra de las proteínas asociadas con una ruta bioquímica de señalización procedente de las muestras de prueba en condiciones que permitirán que se forme un complejo entre el agente y las proteínas asociadas con una ruta bioquímica de señalización. La formación del complejo se puede detectar directa o indirectamente de acuerdo con los procedimientos estándar en la materia. En el método de detección directa, los agentes se suministran con un marcador detectable y los agentes sin reaccionar se pueden eliminar del complejo; la cantidad de marcador restante indica, de este modo, la cantidad de complejo formado. Para dicho método, es preferible seleccionar marcadores que permanezcan unidos a los agentes incluso durante condiciones de lavado rigurosas. Es preferible que, el marcador no interfiera con la reacción de unión. Como alternativa, un procedimiento de detección indirecta puede usar un agente que contenga un marcador introducido químicamente o enzimáticamente. Un marcador deseable, generalmente, no interfiere con la unión o la estabilidad del complejo agente:polipéptido resultante. Sin embargo, el marcador está diseñado, normalmente, para ser accesible a un anticuerpo para una unión eficaz y, por lo tanto, para generar una señal detectable. Se conoce en la materia una amplia variedad de marcadores adecuados para detectar niveles de proteína. Los ejemplos no limitantes incluyen radioisótopos, enzimas, metales coloidales, compuestos fluorescentes, compuestos bioluminiscentes y compuestos quimioluminiscentes.

La cantidad de complejos agente:polipéptido formados durante la reacción de unión se puede cuantificar mediante ensayos cuantitativos estándar. Como se ilustró anteriormente, la formación del complejo agente:polipéptido se puede medir directamente mediante la cantidad de marcador que permanece en el sitio de unión. En una alternativa, la proteína asociada con una ruta bioquímica de señalización se prueba para determinar su capacidad para competir con un análogo marcado para los sitios de unión en el agente específico. En este ensayo competitivo, la cantidad de marcador capturado es inversamente proporcional a la cantidad de secuencias de proteínas asociadas con una ruta bioquímica de señalización presente en una muestra de prueba.

Una serie de técnicas para el análisis de proteínas basadas en los principios generales descritos anteriormente están disponibles en la materia. Incluyen, pero sin limitación, radioinmunoensayos, ELISA (ensayos inmunorradiométricos ligados a enzimas), inmunoensayos en “sándwich”, ensayos de radiometría inmunológica, inmunoensayos in situ (que utilizan, por ejemplo, oro coloidal, marcadores de enzimas o radioisótopos), análisis de transferencia Western, ensayos de precipitación inmunológica, ensayos inmunofluorescentes y SDS-PAGE.

Los anticuerpos que específicamente reconocen o se unen a proteínas asociadas con una ruta bioquímica de

señalización son preferibles para realizar los análisis de proteínas mencionados anteriormente. En los casos en los que se desee, se pueden usar anticuerpos que reconocen un tipo específico de modificaciones postraduccionales (por ejemplo, modificaciones inducibles de la ruta bioquímica de señalización). Las modificaciones postraduccionales incluyen, pero no se limitan a, glicosilación, lipidación, acetilación y fosforilación. Estos anticuerpos pueden comprarse a vendedores comerciales. Por ejemplo, los anticuerpos anti-fosfotirosina que reconocen específicamente las proteínas fosforiladas en tirosina están disponibles en varios proveedores, incluidos Invitrogen y Perkin Elmer. Los anticuerpos anti-fosfotirosina son particularmente útiles en la detección de proteínas que son fosforiladas diferencialmente en sus restos de tirosina en respuesta a un estrés ER. Dichas proteínas incluyen, pero no se limitan a, factor 2 alfa de inicio de la traducción eucariota (eIF-2α). Como alternativa, estos anticuerpos pueden generarse utilizando tecnologías de anticuerpos policlonales o monoclonales convencionales mediante la inmunización de un animal hospedador o una célula productora de anticuerpos con una proteína diana que exhibe la modificación postraduccional deseada.

Puede ser deseable discernir el patrón de expresión de una proteína asociada con una ruta bioquímica de señalización en diferentes tejidos corporales, en diferentes tipos de células y/o en diferentes estructuras subcelulares. Estos estudios se pueden realizar con el uso de anticuerpos específicos de tejido, específicos de la célula o específicos de la estructura subcelular capaces de unirse a marcadores de proteínas que se expresan preferentemente en ciertos tejidos, tipos de células o estructuras subcelulares.

Una expresión alterada de un gen asociado con una ruta bioquímica de señalización también puede determinarse mediante el examen de un cambio en la actividad del producto génico en relación con una célula de control. El ensayo de un cambio inducido por el agente en la actividad de una proteína asociada con una ruta bioquímica de señalización dependerá de la actividad biológica y/o la ruta de transducción de señales que se esté investigando. Por ejemplo, cuando la proteína es una quinasa, se puede determinar un cambio en su capacidad para fosforilar el sustrato(s) cadena abajo mediante varios ensayos conocidos en la materia. Los ensayos representativos incluyen, pero no se limitan a, inmunotransferencia e inmunoprecipitación con anticuerpos tales como anticuerpos anti-fosfotirosina que reconocen proteínas fosforiladas. Además, la actividad de la quinasa se puede detectar mediante ensayos quimioluminiscentes de alto rendimiento, tales como AlphaScreen™ (disponible en Perkin Elmer) y el ensayo eTag™ (Chan-Hui, et al. (2003) Clinical Immunology 111: 162-174).

Cuando la proteína asociada con una ruta bioquímica de señalización es parte de una cascada de señalización que conduce a una fluctuación de la condición de pH intracelular, las moléculas sensibles al pH, tales como los colorantes de pH fluorescentes, pueden usarse como moléculas indicadoras. En otro ejemplo en el que la proteína asociada con una ruta bioquímica de señalización es un canal iónico, se pueden monitorear las fluctuaciones en el potencial de membrana y/o la concentración de iones intracelular. Un número de kits comerciales y dispositivos de alto rendimiento son particularmente adecuados para una detección rápida y robusta de moduladores de canales iónicos. Los instrumentos representativos incluyen FLIPRTM (Molecular Devices, Inc.) y VIPR (Aurora Biosciences). Estos instrumentos son capaces de detectar reacciones en más de 1000 pocillos de muestra de una microplaca simultáneamente, y proporcionar mediciones en tiempo real y datos funcionales en un segundo o incluso en un minisegundo.

En la práctica de cualquiera de los métodos divulgados en el presente documento, se puede introducir un vector adecuado en una célula o un embrión a través de uno o más métodos conocidos en la materia, incluyendo sin limitación, microinyección, electroporación, sonoporación, biolistas, transfección mediada por fosfato de calcio, transfección catiónica, transfección de liposomas, transfección de dendrímeros, transfección de choque térmico, transfección de nucleofección, magnetofeccion, lipofección, impalefeccion, transfección óptica, absorción de ácidos nucleicos patentada por el agente y el suministro a través de liposomas, inmunoliposomas, virosomas, o viriones artificiales. En algunos métodos, el vector se introduce en un embrión mediante microinyección. El vector o vectores pueden ser microinyectados en el núcleo o el citoplasma del embrión. En algunos métodos, el vector o vectores pueden introducirse en una célula mediante nucleofección.

El polinucleótido diana de un complejo CRISPR puede ser cualquier polinucleótido endógeno o exógeno a la célula eucariota. Por ejemplo, el polinucleótido diana puede ser un polinucleótido que reside en el núcleo de la célula eucariota. El polinucleótido diana puede ser una secuencia que codifica un producto génico (por ejemplo, una proteína) o una secuencia no codificante (por ejemplo, un polinucleótido regulador o un ADN basura).

Los ejemplos de polinucleótidos diana incluyen una secuencia asociada con una ruta bioquímica de señalización, por ejemplo, un gen o polinucleótido asociado con la ruta bioquímica de señalización. Los ejemplos de polinucleótidos diana incluyen un gen o polinucleótido asociado a enfermedad. Un gen o polinucleótido "asociado a enfermedad" se refiere a cualquier gen o polinucleótido que produce productos de transcripción o traducción a un nivel anormal o en una forma anormal en células procedentes de tejidos afectados por una enfermedad en comparación con tejidos o células de un control que no controla la enfermedad. Puede ser un gen que se expresa en un nivel anormalmente alto; puede ser un gen que se expresa en un nivel anormalmente bajo, donde la expresión alterada se correlaciona con la aparición y/o la progresión de la enfermedad. Un gen asociado a enfermedad también se refiere a un gen que posee mutaciones o variaciones genéticas que son directamente responsables o están en desequilibrio de enlace con un gen(es) que es responsable de la etiología de una enfermedad. Los productos transcritos o traducidos pueden ser

conocidos o desconocidos, y pueden estar en un nivel normal o anormal.

El polinucleótido diana de un complejo CRISPR puede ser cualquier polinucleótido endógeno o exógeno a la célula eucariota. Por ejemplo, el polinucleótido diana puede ser un polinucleótido que reside en el núcleo de la célula eucariota. El polinucleótido diana puede ser una secuencia que codifica un producto génico (por ejemplo, una proteína) o una secuencia no codificante (por ejemplo, un polinucleótido regulador o un ADN basura),

El polinucleótido objetivo de un complejo CRISPR puede incluir una cantidad de genes y polinucleótidos asociados a enfermedades, así como los genes y polinucleótidos asociados a rutas bioquímicas de señalización, como se enumeran en las solicitudes de patentes provisionales de Estados Unidos 61/736,527 y 61/748,427 que tienen la referencia general BI-2001/008/WSGR Docket N.º 44063-701.101 y BI-2011/008/WSGR Docket N.º 44063-701.102 respectivamente, ambos titulados SYSTEMS METHODS AND COMPOSITIONS FOR SEQUENCE MANIPULATION presentados el 12 de diciembre de 2012 y el 2 de enero de 2013, respectivamente.

Los ejemplos de polinucleótidos diana incluyen una secuencia asociada con una ruta bioquímica de señalización, por ejemplo, un gen o polinucleótido asociado con la ruta bioquímica de señalización. Los ejemplos de polinucleótidos diana incluyen un gen o polinucleótido asociado a enfermedad. Un gen o polinucleótido "asociado a enfermedad" se refiere a cualquier gen o polinucleótido que produce productos de transcripción o traducción a un nivel anormal o en una forma anormal en células procedentes de tejidos afectados por una enfermedad en comparación con tejidos o células de un control que no controla la enfermedad. Puede ser un gen que se expresa en un nivel anormalmente alto; puede ser un gen que se expresa en un nivel anormalmente bajo, donde la expresión alterada se correlaciona con la aparición y/o la progresión de la enfermedad. Un gen asociado a enfermedad también se refiere a un gen que posee mutaciones o variaciones genéticas que son directamente responsables o están en desequilibrio de enlace con un gen(es) que es responsable de la etiología de una enfermedad. Los productos transcritos o traducidos pueden ser conocidos o desconocidos, y pueden estar en un nivel normal o anormal.

Las realizaciones también se relacionan con métodos y composiciones relacionadas con la desactivación de genes, la amplificación de genes y la reparación de mutaciones particulares asociadas con la inestabilidad de la repetición de ADN y los trastornos neurológicos (Robert D, Wells, Tetsuo Ashizawa, Genetic Instabilities and Neurological Diseases, segunda edición, Academic Press, Oct 13, 2011 - Medical). Se ha encontrado que los aspectos específicos de las secuencias de repetición en tándem son responsables de más de veinte enfermedades humanas (New insights into repeat instability: role of RNA•DNA hybrids, Mclvor El, Polak U, Napierala M. RNA Biol. 2010 Sep-Oct;7(5):551-8). El sistema CRISPR-Cas puede aprovecharse para corregir estos defectos de inestabilidad genómica. Y por lo tanto, las secuencias diana se pueden encontrar en estos defectos de inestabilidad genómica.

Otras realizaciones se relacionan con algoritmos que establecen la base de los métodos relacionados con la enzima CRISPR, por ejemplo, Cas, especificidad o actividad inespecífica. En general, los algoritmos se refieren a un método eficaz expresado como una lista finita de instrucciones bien definidas para calcular una o más funciones de interés. Los algoritmos pueden expresarse en varios tipos de notación, incluidos, pero sin limitación, lenguajes de programación, diagramas de flujo, tablas de control, lenguajes naturales, fórmulas matemáticas y pseudocódigos. En una realización preferida, el algoritmo puede expresarse en un lenguaje de programación que expresa el algoritmo en una forma que puede ejecutarse mediante una computadora o un sistema informático,

Los métodos relacionados con la enzima CRISPR, por ejemplo, Cas, la especificidad o la actividad inespecífica se basan en algoritmos que incluyen, pero sin limitación, el algoritmo termodinámico, el algoritmo multiplicativo y el algoritmo posicional. Estos algoritmos toman una entrada de una secuencia de interés e identifican secuencias diana candidatas para luego proporcionar una salida de una clasificación de secuencias diana candidatas o una puntuación asociada con una secuencia diana particular basada en sitios predichos inespecíficos. Los sitios diana candidatos pueden seleccionarse mediante un usuario final o un cliente en función de consideraciones que incluyen, pero no se limitan a, la eficacia de la modificación, el número o la ubicación de la escisión inespecífica prevista. En una realización más preferida, un sitio diana candidato es único o tiene una escisión pronosticada mínima inespecífica dados los parámetros anteriores. Sin embargo, la relevancia funcional de la posible modificación inespecífica también debe considerarse cuando se elige un sitio diana, en particular, un usuario final o un cliente puede considerar si los sitios inespecíficos ocurren dentro de loci de función genética conocida, es decir, exones que codifican proteínas, regiones potenciadoras o elementos reguladores intergénicos. También puede haber consideraciones específicas del tipo de célula, es decir, si se produce un sitio inespecífico en un lugar que no es funcionalmente relevante en el tipo de célula diana. En conjunto, un usuario final o cliente puede realizar luego una selección informada y específica de la aplicación de un sitio diana candidato con una modificación mínima inespecífica,

El algoritmo termodinámico se puede aplicar mediante la selección de un complejo CRISPR para la orientación y/o escisión de una secuencia de ácido nucleico diana candidata en una célula. El primer paso es ingresar la secuencia diana (Paso S400) que puede haber sido determinada utilizando el algoritmo de posición. También se introduce un complejo CRISPR (Paso S402). El siguiente paso es comparar la secuencia diana con la secuencia guía para el complejo CRISPR (Paso S404) para identificar cualquier desapareamiento. Además, se puede determinar la cantidad, ubicación y naturaleza de los desapareamientos entre la secuencia guía del potencial complejo CRISPR y la secuencia de ácido nucleico diana candidata. Luego se calcula la energía libre de hibridación de la unión entre la secuencia diana

y la secuencia guía (Paso S406). Por ejemplo, esto puede calcularse determinando una contribución de cada una de la cantidad, ubicación y naturaleza de los desapareamientos en la energía libre de hibridación de la unión entre la secuencia de ácido nucleico diana y la secuencia guía de los complejos CRISPR potenciales. Además, esto se puede calcular aplicando un modelo calculado que utiliza un conjunto de datos de capacitación como se explica con más detalle a continuación. Sobre la base de la energía libre de hibridación (es decir, en función del análisis de contribución), se genera una predicción de la probabilidad de escisión en la ubicación(es) del desapareamiento(s) de la secuencia de ácido nucleico diana por el potencial de complejo(s) CRISPR (Paso S408). Luego, el sistema determina si hay o no complejos CRISPR adicionales a considerar y, si es así, repite los pasos de comparación, cálculo y predicción. Cada complejo CRISPR se selecciona del posible complejo(s) CRISPR en función de si la predicción indica que es más probable que la escisión no se produzca en la ubicación(es) de desapareamiento(s) mediante el complejo CRISPR (Paso S410). Opcionalmente, las probabilidades de escisión pueden clasificarse de modo que se seleccione un complejo CRISPR único. La determinación de la contribución de cada uno de la cantidad, ubicación y naturaleza de los desapareamientos en la energía libre de hibridación incluye, pero sin limitación, determinar la contribución relativa de estos factores. El término "ubicación", tal como se utiliza en el término "ubicación de desapareamiento(s)" puede referirse a la ubicación real de uno o más desapareamientos de pares de bases, pero también puede incluir la ubicación de un tramo de pares de bases que flanquean el desapareamiento(s) de pares de bases o un intervalo de ubicaciones/posiciones. El tramo de pares de bases que flanquean los desapareamientos de pares de bases puede incluir, pero sin limitaciones, al menos uno, al menos dos, al menos tres pares de bases, al menos cuatro o al menos, cinco o más pares de bases a cada lado de uno o más desapareamiento(s). Como se usa en el presente documento, la "energía libre de hibridación" puede ser una estimación de la energía libre de unión, por ejemplo, energía de unión libre de ADN:ARN que se puede estimar a partir de datos sobre energía libre de unión de ADN:ARN y energía libre de unión de ARN:ARN.

En los métodos relacionados con el algoritmo multiplicativo aplicado en la identificación de una o más secuencias diana únicas en un genoma de un organismo eucariota, mediante el cual la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) crear un conjunto de datos de capacitación en cuanto a una Cas particular, b) determinar la frecuencia de corte promedio en una posición particular para la Cas particular a partir del conjunto de datos de capacitación, c) determinar la frecuencia de corte promedio de un desapareamiento particular para la Cas particular a partir del conjunto de datos de capacitación, d) multiplicar la frecuencia de corte promedio en una posición particular mediante la frecuencia de corte promedio de un desapareamiento particular para obtener un primer producto, e) repetir los pasos b) a d) para obtener un segundo y más productos para cualquier otra posición(es) particular(s) de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para obtener un producto final, y omitir este paso si no hay desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular (u opcionalmente e) repetir los pasos b) a d) para obtener un segundo y más productos para cualquier posición(es) particular adicional de desapareamientos y desapareamientos particulares y multiplicar estos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento en particular en una posición particular), y f) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por 18 y omitir este paso si no hay un desapareamiento en ninguna posición o si solo hay un desapareamiento particular en una posición particular (u opcionalmente f) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos entre 18 y omitir este paso si no hay un desapareamiento en ninguna posición o si hay solo un desapareamiento particular en una posición particular), para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas, las frecuencias de corte pronosticadas para dianas de todo el genoma se pueden calcular mediante la multiplicación, en serie: $f_{sst} = f(1)g(N_1,N'_1)\times f(2)g(N_2,N'_2)\times\cdots f(19)g(N_{19},N'_{19})\times h$ con valores $f(i)$ y $g(N_t,N'_t)$ en la posición t correspondiente, respectivamente, a las frecuencias de corte de desapareamientos de posición agregada y de base para posiciones y emparejamientos indicados en una matriz de transición de base generalizada o una matriz de agregado, por ejemplo, una matriz como se indica en la Figura 12c. Cada frecuencia se normalizó en un intervalo de 0 a 1, de manera que $f\rightarrow(f-f_{min})/(f_{max}-f_{min})$. En el caso de una coincidencia, ambos se establecieron igual a 1. Mientras tanto, el valor h volvió a ponderar la frecuencia estimada mediante la distancia por pares mínima entre desapareamientos consecutivos en la secuencia objetivo. Esta distancia del valor, en pares de bases, se dividió entre 18 para dar un valor máximo de 1 (en los casos en que existían menos de 2 desapareamientos, o cuando los desapareamientos ocurrían en los extremos opuestos del margen diana de 19 pb). Las muestras que tienen un recuento de lecturas de al menos 10.000 (n = 43) se representaron gráficamente. A los empatados en la clasificación se les dio un promedio de clasificación. El coeficiente de correlación de Spearman, 0,58, indicó que las frecuencias estimadas recapitularon el 58 % de la varianza de clasificación para las frecuencias de corte observadas. La comparación de f_{sst} con las frecuencias de corte produjo directamente una correlación de Pearson de 0,89. Si bien está dominado por los pares ARNg/diana de mayor frecuencia, este valor indica que casi el 90 % de toda la varianza de la frecuencia de corte se explica por las predicciones anteriores. En aspectos adicionales de la divulgación, el algoritmo multiplicativo o los métodos mencionados en el presente documento también pueden incluir factores termodinámicos, por ejemplo, energías de hibridación u otros factores de interés se multiplican en serie para llegar al producto final.

En realizaciones de la divulgación, la determinación de la actividad inespecífica de una enzima CRISPR puede permitir a un usuario final o un cliente predecir los mejores sitios de corte en un lugar de interés genómico. En una realización adicional de la invención, se puede obtener una clasificación de las frecuencias de corte en varios sitios supuestos

inespecíficos para verificar *in vitro*, *in vivo* o *ex vivo* si uno o más de los peores escenarios de corte inespecífico hacen o no hacen que ocurra. En otra realización de la divulgación, la determinación de la actividad inespecífica puede ayudar en la selección de sitios específicos si un usuario final o cliente está interesado en maximizar la diferencia entre la frecuencia de corte en la diana y la frecuencia de corte más alta obtenida en la clasificación de los sitios inespecíficos. Otro aspecto de la selección incluye revisar la clasificación de los sitios e identificar los loci genéticos de las dianas no específicos para asegurar que un sitio diana específico seleccionado tenga la diferencia apropiada en la frecuencia de corte de las diana que pueden codificar para oncogenes u otros loci genéticos de interés. Los aspectos de la divulgación pueden incluir métodos para minimizar el riesgo terapéutico mediante la verificación de la actividad inespecífica del complejo CRISPR-Cas. Otros aspectos de la divulgación pueden incluir el uso de información sobre la actividad inespecífica del complejo CRSIPR-Cas para crear sistemas de modelos específicos (por ejemplo, un ratón) y líneas celulares. Los métodos de la invención permiten un análisis rápido de efectos no específicos y pueden aumentar la eficacia de un laboratorio.

En los métodos relacionados con el algoritmo de posición aplicado en la identificación de una o más secuencias diana únicas en un genoma de un organismo eucariota, mediante el cual la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende: a) determinar la frecuencia de corte promedio de los desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación en cuanto a esa Cas, si hay más de un desapareamiento, repita el paso a) para determinar la frecuencia de corte para cada desapareamiento, multiplicar las frecuencias de los desapareamientos para obtener de ese modo una clasificación, que permite la identificación de una o más secuencias diana únicas, se puede ver un ejemplo de una aplicación de este algoritmo en la Fig. 23.

La figura 32, 33A, 33B y 34, respectivamente, muestran un diagrama de flujo de los métodos de la invención. La Figura 32 proporciona un diagrama de flujo con respecto a los métodos de ubicación o de posición de la invención, es decir, con respecto a la identificación computacional de sitios diana CRISPR únicos: Para identificar sitios diana únicos para una Cas, por ejemplo, una Cas9, por ejemplo, la enzima SF370 Cas9 de S. pyogenes (SpCas9), en moléculas de ácido nucleico, por ejemplo, de células, por ejemplo, de organismos, que incluyen, pero no se limitan a, seres humanos, ratón, rata, pez cebra, mosca de la fruta y genoma de C. elegans, los solicitantes desarrollaron un paquete de software para escanear ambas cadenas de una secuencia de ADN e identificar todos los posibles sitios diana de SpCas9. El método se muestra en la Figura 32, que muestra que el primer paso es ingresar la secuencia del genoma (Paso S100). Se seleccionan el motivo(s) CRISPR que son adecuados para esta secuencia del genoma (Paso S102). En cuanto a este ejemplo, el motivo CRISPR es una secuencia de motivo adyacente al protoespaciador (PAM) NGG. Luego se selecciona un fragmento de longitud fija que debe ocurrir en la secuencia general antes del motivo seleccionado (es decir, cadena arriba en la secuencia) (Paso S102). En este caso, el fragmento es una secuencia de 20 pb. Por lo tanto, cada sitio diana de SpCas9 se definió operativamente como una secuencia de 20 pb seguida de una secuencia de motivo adyacente al protoespaciador (PAM) NGG, y se identificaron todas las secuencias que satisfacen esta definición 5'-N20-NGG-3' en todos los cromosomas (Paso S106). Para evitar la edición no específica del genoma, después de identificar todos los sitios potenciales, todos los sitios diana se filtraron en función del número de veces que aparecen en el genoma de referencia relevante (Paso S108). (Esencialmente, se agregan todos los fragmentos de 20 pb (sitios diana candidatos) cadena arriba del motivo PAM de NGG). Si un fragmento de 20 pb en particular aparece más de una vez en la búsqueda de todo el genoma, se considera que no es único y se "elimina", también conocido como filtrado. Los fragmentos de 20 pb que permanecen, por lo tanto, aparecen solo una vez en el genoma diana, lo que los hace únicos; y, en lugar de tomar un fragmento de 20 bp (el sitio diana de Cas9 completo), este algoritmo toma el primero, por ejemplo, ll-12 pb cadena arriba del motivo PAM y requiere que sea único.) Por último, se selecciona un sitio diana único (Paso S110), por ejemplo, para aprovechar la especificidad de secuencia de Cas, por ejemplo, la actividad de Cas9 conferida por una secuencia 'semilla', que pueden ser, por ejemplo, aproximadamente la secuencia ll-12 pb en 5' de la secuencia PAM, se seleccionaron las secuencias 5'-NNNNNNNNNN-NGG-3' para que sean únicas en el genoma relevante. Las secuencias genómicas están disponibles en UCSC Genome Browser y en las visualizaciones de muestra de la información del genoma humano hg, genoma de ratón mm, genoma de rata rn, genoma de pez cebra danRer, genoma de D. melanogaster dm, genoma de C. elegans ce, el genoma porcino y el genoma de vaca se muestran en las figura 15 a 22 respectivamente.

Las figuras 33A y 33B proporcionan cada una un diagrama de flujo en cuanto a los métodos termodinámicos de la invención. La Figura 34 proporciona un diagrama de flujo de los métodos de multiplicación de la invención. Con referencia a las Figuras 33A y 33B, y considerando el modelo termodinámico de mínimos cuadrados de la eficacia de corte de CRISPR-Cas, para sitios diana de Cas9 arbitrarios, los solicitantes generaron un modelo termodinámico numérico que predice la eficacia de corte de Cas9. Los solicitantes proponen 1) que el ARN guía de Cas9 tiene energías libres específicas de hibridación a su diana y cualquier secuencia de ADN inespecífica y 2) que Cas9 modifica las energías libres de hibridación de ARN:ADN localmente de una manera dependiente de la posición pero independiente de la secuencia. Los solicitantes captaron un modelo para predecir la eficacia de corte de CRISPR-Cas basándose en sus datos de mutación de ARN guía de CRISPR-Cas y en los cálculos de energía termodinámica ARN:ADN utilizando un algoritmo de aprendizaje automático. Luego, los solicitantes validaron sus modelos resultantes mediante la comparación de sus predicciones de corte inespecíficos de CRISPR-Cas en múltiples loci genómicos con datos experimentales que evalúan la modificación de locus en los mismos sitios. La metodología adoptada en el desarrollo de este algoritmo es la siguiente: El resumen del problema indica que para espaciadores arbitrarios y dianas de longitud constante, se debe encontrar un modelo numérico que tenga sentido termodinámico y prediga la eficacia

de corte de Cas9. Suponiendo que Cas9 modifica las energías libres de hibridación de ADN:ARN localmente de forma dependiente de la posición pero independiente de la secuencia. El primer paso es definir un modelo que tenga un conjunto de pesos que vincule la energía libre de hibridación Z con las energías libres locales *G* (Paso S200). Luego, para las energías libres de hibridación de ADN:ARN $\Delta G_{ij}(k)$ (para la posición *k* entre 1 y N) del espaciador *i* y de la diana *j*

$$Z_{ij} = \sum_{k=1}^{N} \alpha_k \Delta G_{ij}(k)$$

Z_{ij} puede tratarse como una energía libre "eficaz" modificada por la posición multiplicativa-pesos α_k. El Z_{ij} "eficaz" de energía libre corresponde a una probabilidad de corte asociada ~ $e^{-\beta Z_{ij}}$ (para algunos β constantes) de la misma manera que un modelo de equilibrio de hibridación (sin ponderación de posición) habría predicho una probabilidad de hibridación ~$e^{-\beta \Delta G_{ij}}$. Dado que se ha medido la eficacia de corte, los valores Z_{ij} pueden tratarse como sus observables. Paralelamente, se puede calcular $\Delta G_{ij}(k)$ para el emparejamiento espaciador-diana de cualquier experimento. La tarea de los solicitantes era encontrar los valores α_k, ya que esto les permitiría estimar Z_{ij} para cualquier par espaciador-diana. Los pesos se determinan mediante el ingreso de valores conocidos para Z y G a partir de un conjunto de secuencias de capacitación con los valores conocidos que se determinan mediante la experimentación según sea necesario. Por lo tanto, los solicitantes deben definir un conjunto de secuencias de capacitación (Paso S202) y calcular un valor de Z para cada secuencia en el conjunto de capacitación (Paso S204). Escribiendo la ecuación anterior para Z_{ij} en forma matricial, los solicitantes obtienen:

$$\vec{Z} = \mathbf{G}\vec{\alpha} \qquad (1)$$

La estimación de mínimos cuadrados es entonces

$$\vec{\alpha}_{est} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\vec{Z}$$

donde G^T es la matriz-transposición de la *G* y $(G^TG)^{-1}$ es la inversa de su matriz-producto, En la anterior, **G** es una matriz de valores locales de energía libre de ADN:ARN cuya fila *r* corresponde al ensayo experimental *r* y cuya columna *k* corresponde a la posición *k* en el híbrido de ADN:ARN probado en ese ensayo experimental. Estos valores de G se ingresan en el sistema de capacitación (Paso S204). Mientras tanto, Z es un vector-columna cuya fila *r* corresponde a los observables del mismo ensayo experimental que la fila *r* de **G**. Debido a la relación descrita anteriormente en la que se estima que las frecuencias de corte de CRISPR varían según ~ $e^{-\beta Z_{ij}}$, estos observables, Z_{ij}, se calcularon como el logaritmo natural de la frecuencia de corte observada. Lo observable es la eficacia de escisión de Cas, por ejemplo, Cas9, en un ADN diana para un ARN guía particular y un par de ADN dianas. El experimento es Cas, por ejemplo, Cas9, con un emparejamiento ARNsg/ADN diana particular, y el observable es el porcentaje de escisión (ya sea medido como porcentaje de formación de indel de células o simplemente porcentaje de escisión *in vitro*) (véase en el presente documento la discusión sobre cómo generar un conjunto de datos de capacitación). Más en particular, cada reacción de PCR única que fue secuenciada debe tratarse como un ensayo experimental único para abarcar la replicabilidad dentro del vector. Esto significa que cada réplica experimental va en filas separadas de la ecuación 1 (y debido a esto, algunas filas de **G** serán idénticas). La ventaja de esto es que cuando $\vec{\alpha}$ se ajusta, toda la información relevante, incluida la replicabilidad, se tiene en cuenta en la estimación final. Los valores $\vec{Z}$ observables, se calcularon como log (frecuencia de corte observada) (Paso S206). Las frecuencias de corte se normalizaron opcionalmente de manera idéntica (de modo que todas tienen las mismas "unidades") (Paso S208). Sin embargo, para conectar valores de frecuencia indel de secuencia, puede ser mejor estandarizar la profundidad de secuenciación. La forma preferida de hacer esto sería establecer una profundidad de secuencia estándar *D* para la cual todos los experimentos incluidos en $\vec{Z}$ tengan al menos ese número de lecturas. Debido a que las frecuencias de corte por debajo de 1/*D* no se pueden detectar de manera consistente, debe establecerse como la frecuencia mínima para el conjunto de datos, y los valores en $\vec{Z}$ deben variar desde log(1/*D*) a log(1). Se podría variar el valor de *D* más adelante para asegurarse de que la estimación de $\vec{\alpha}$ no sea demasiado dependiente del valor elegido. Por lo tanto, los valores de Z podrían filtrarse si no alcanzan la profundidad de secuencia mínima (Paso S210). Una vez que los valores de G y Z se ingresan en el sistema de aprendizaje automático, se pueden determinar los pesos (Paso S212) y los resultados (Paso S214). Estos pesos se pueden usar para estimar la energía libre Z y la frecuencia de corte para cualquier secuencia. En un aspecto adicional, existen diferentes métodos para graficar secuencias NGG y NNAGAAW. Uno es con el método 'no superpuesto'. NGG y NRG pueden registrarse de forma "superpuesta", como se indica en las figuras 6A-C. Los solicitantes también realizaron un estudio sobre la actividad de Cas9 inespecífica como se indica en las figuras 10, 11 y 12. Los aspectos de la divulgación también se relacionan con modelos predictivos que pueden no involucrar energías de hibridación, sino que simplemente usan la información de frecuencia de corte como una predicción.

La figura 34 muestra los pasos en un método relacionado con el algoritmo multiplicativo que puede aplicarse para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, mediante el cual la secuencia

diana es susceptible de ser reconocida por un sistema CRISPR-Cas. El método comprende: a) crear un conjunto de datos de capacitación como para una Cas particular. El conjunto de datos de capacitación se puede crear como se describe con más detalle más adelante mediante la determinación de los pesos asociados con un modelo. Una vez que se ha establecido un conjunto de datos de capacitación, se puede usar para predecir el comportamiento de una secuencia de entrada e identificar una o más secuencias diana únicas en el mismo. En el paso S300, la secuencia del genoma se introduce en el sistema. En una Cas particular, el siguiente paso es ubicar un desapareamiento entre una secuencia diana dentro de la secuencia de entrada y el ARN guía para la Cas en particular (Paso S302). Para el desapareamiento identificado, se determinan dos frecuencias de corte promedio utilizando el conjunto de datos de capacitación. Estos son la frecuencia de corte promedio en la posición del desapareamiento (paso S304) y la frecuencia de corte promedio asociada con ese tipo de desapareamiento (Paso S306). Estas frecuencias de corte promedio se determinan a partir del conjunto de datos de capacitación que es específico de esa Cas. El siguiente paso S308 es crear un producto multiplicando la frecuencia de corte promedio en una posición particular por la frecuencia de corte promedio de un desapareamiento particular para obtener un primer producto. Luego se determina en el paso S310 si hay o no otros desapareamientos. Si no hay ninguno, la secuencia diana se genera como la secuencia diana única. Sin embargo, si hay otros desapareamientos, los pasos 304 a 308 se repiten para obtener un segundo y más productos para cualquier posición(es) particular adicional de desapareamientos y desapareamientos particulares. Donde se crean segundos y más productos y todos los productos se multiplican para crear un producto final. Luego, el producto final se multiplica por el resultado de dividir la distancia mínima entre desapareamientos consecutivos entre la longitud de la secuencia diana (por ejemplo, 18) (paso S314) que escala con eficacia cada producto final. Se apreciará que los pasos 312 y 314 se omiten si no hay un desapareamiento en ninguna posición o si hay solo un desapareamiento particular en una posición particular. El proceso se repite luego para cualquier otra secuencia diana. Los productos finales "a escala" para cada secuencia diana se clasifican para obtener así una clasificación (Paso S316), que permite la identificación de una o más secuencias diana únicas mediante la selección de la más alta clasificada (Paso S318). Por lo tanto, el producto final "a escala" que representa las frecuencias de corte predichas para dianas de todo el genoma se puede calcular de la siguiente manera: $f_{sst} = f(1)g(N_1,N'_1)\times f(2)g(N_2,N'_2)\times\cdots f(19)g(N_{19},N'_{19})\times h$

con valores $f(i)$ y $g(N_i,N'_i)$ en la posición i correspondiente, respectivamente, a las frecuencias de corte de desapareamientos de posición agregada y de base para posiciones y emparejamientos indicados en una matriz de transición de base generalizada o una matriz de agregado, por ejemplo, una matriz como se indica en la Figura 12c. En otras palabras, f(i) es la frecuencia de corte promedio en la posición particular para el desapareamiento y $g(N_i,N'_i)$ es la frecuencia de corte promedio para el tipo de desapareamiento particular para el desapareamiento. Cada frecuencia se normalizó en un intervalo de 0 a 1, de manera que $f\rightarrow(f-f_{min})/(f_{max}-f_{min})$. En el caso de una coincidencia, ambos se establecieron igual a 1, Mientras tanto, el valor h volvió a ponderar la frecuencia estimada mediante la distancia por pares mínima entre desapareamientos consecutivos en la secuencia diana. Esta distancia del valor, en pares de bases, se dividió entre una constante que era indicativa de la longitud de la secuencia diana (por ejemplo, 18) para dar un valor máximo de 1 (en los casos en que existían menos de 2 desapareamientos, o cuando los desapareamientos ocurrían en los extremos opuestos del margen diana de 19 pb). Las muestras que tienen un recuento de lecturas de al menos 10.000 (n = 43) se representaron gráficamente. A los empatados en la clasificación se les dio un promedio de clasificación. El coeficiente de correlación de Spearman, 0,58, indicó que las frecuencias estimadas recapitularon el 58 % de la varianza de clasificación para las frecuencias de corte observadas. La comparación de f_{sst} con las frecuencias de corte produjo directamente una correlación de Pearson de 0,89. Si bien está dominado por los pares ARNg/diana de mayor frecuencia, este valor indica que casi el 90 % de toda la varianza de la frecuencia de corte se explica por las predicciones anteriores. En aspectos adicionales de la divulgación, el algoritmo multiplicativo o los métodos mencionados en el presente documento también pueden incluir factores termodinámicos, por ejemplo, energías de hibridación u otros factores de interés se multiplican en serie para llegar al producto final.

La figura 35 muestra un diagrama de bloques esquemático de un sistema informático que se puede utilizar para implementar los métodos descritos en el presente documento. El sistema informático 50 comprende un procesador 52 acoplado a la memoria de código y datos 54 y un sistema de entrada/salida 56 (por ejemplo, que comprende interfaces para una red y/o medios de almacenamiento y/u otras comunicaciones). El código y/o los datos almacenados en la memoria 54 pueden proporcionarse en un medio de almacenamiento extraíble 60. También puede haber una interfaz 58 de usuario que comprenda, por ejemplo, un teclado y/o un ratón y una pantalla 62 de usuario. El sistema informático está conectado a una base de datos 78. La base de datos 78 comprende los datos asociados con los conjuntos de datos de capacitación. El sistema informático se muestra como un único dispositivo informático con múltiples componentes internos que pueden implementarse desde una sola o varias unidades centrales de procesamiento, por ejemplo, microprocesadores. Se apreciará que la funcionalidad del dispositivo puede distribuirse en varios dispositivos informáticos. También se apreciará que los componentes individuales pueden combinarse en uno o más componentes que proporcionan la funcionalidad combinada. Además, cualquiera de los módulos, bases de datos o dispositivos mostrados pueden implementarse en una computadora de propósito general modificada (por ejemplo, programada o configurada) por software para que sea una computadora de propósito especial para realizar las funciones descritas en el presente documento. El procesador puede estar configurado para llevar a cabo los pasos que se muestran en los distintos diagramas de flujo. La interfaz de usuario se puede usar para ingresar la secuencia del genoma, el motivo CRISPR y/o Cas para los cuales se debe identificar una secuencia diana. Las secuencias diana únicas de salida pueden mostrarse en la pantalla del usuario.

Ejemplos

Los siguientes ejemplos se dan con el propósito de ilustrar varias realizaciones de la invención y no pretenden limitar la presente invención de ninguna manera. Los presentes ejemplos, junto con los métodos descritos en el presente documento son actualmente representativos de realizaciones preferidas, son ejemplos y no pretenden ser limitaciones en el alcance de la invención.

Ejemplo 1: Evaluación de la especificidad de la escisión del genoma mediada por Cas9.

Los solicitantes llevaron a cabo una prueba inicial para evaluar la especificidad de escisión de Cas9 de *Streptococcus pyogenes*. El ensayo se diseñó para probar el efecto de los desapareamientos de pares de bases individuales entre la secuencia de ARN guía y el ADN diana. Los resultados de la ronda inicial de pruebas se muestran en la figura 3.

Los solicitantes llevaron a cabo el ensayo utilizando células 293FT en placas de 96 pocillos. Se transfectaron las células con 65 ng de un plásmido que llevaba Cas9 y 10 ng de un amplicón de PCR que llevaba el promotor U6 pol3 y el ARN guía. El experimento se realizó utilizando una gran cantidad de Cas9 y ARN guía, lo que probablemente explica la especificidad aparentemente baja (es decir, los desapareamientos de una sola base no son suficientes para abolir la escisión). Los solicitantes también evalúan el efecto de diferentes concentraciones de Cas9 y ARN en la especificidad de la escisión. Además, los solicitantes llevan a cabo una evaluación exhaustiva de cada posible desapareamiento en cada posición del ARN guía. El objetivo final es generar un modelo para informar el diseño de los ARN guía que tienen alta especificidad de escisión.

Experimentos adicionales prueban la posición y el número de desapareamientos en el ARN guía en la eficacia de escisión. La siguiente tabla muestra una lista de 48 posibilidades de desapareamiento. En la tabla, 0 significa sin mutación y 1 significa con mutación.

	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	NGG
Regla 1 de la prueba: Más desapareamientos= mayor efecto sobre el corte																					
Regla 2 de la prueba: Los desapareamientos en el extremo 5´ tienen menos efecto que los desapareamientos en el extremo 3´																					
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
8	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	
14	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	
15	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
21	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Regla 3 de la prueba: Los desapareamientos más distribuidos tienen menos efecto que los desapareamientos más concentrados																					
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	
25	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
26	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
27	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	
29	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	
30	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
31	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
32	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	
33	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
34	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	
35	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	
37	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	
38	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	
39	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	
40	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	
41	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	
42	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	
43	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	
44	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	
45	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	
46	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	
47	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	
48	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	

Ejemplo 2: Evaluación de mutaciones en la secuencia PAM, y su efecto sobre la eficacia de escisión.

Los solicitantes probaron mutaciones en la secuencia PAM y su efecto sobre la escisión. La secuencia PAM para Cas9 de *Streptococcus pyogenes* es NGG, donde se cree que GG se requiere para la escisión. Para probar si Cas9 puede escindir secuencias con PAM que son diferentes de NGG, los solicitantes eligieron los siguientes 30 sitios diana del locus Emxl del genoma humano - 2 para cada una de las 15 posibilidades de PAM: NAA, NAC, NAT, NAG, NCA, NCC, NCG, NCT, NTA, NTC, NTG, NTT, NGA, NGC y NGT; NGG no se selecciona porque se puede dirigir de manera eficaz.

Los datos de eficacia de escisión se muestran en la figura 4. Los datos muestran que, aparte de NGG, solo las secuencias con PAM NAG pueden ser dirigidas.

PAM	Diana 1	Diana 2
NAA	AGGCCCCAGTGGCTGCTCT	TCATCTGTGCCCCTCCCTC
NAT	ACATCAACCGGTGGCGCAT	GGGAGGACATCGATGTCAC
NAC	AAGGTGTGGTTCCAGAACC	CAAACGGCAGAAGCTGGAG
NAG	CCATCACATCAACCGGTGG	GGGTGGGCAACCACAAACC
NTA	AAACGGCAGAAGCTGGAGG	GGTGGGCAACCACAAACCC
NTT	GGCAGAAGCTGGAGGAGGA	GGCTCCCATCACATCAACC
NTC	GGTGTGGTTCCAGAACCGG	GAAGGGCCTGAGTCCGAGC
NTG	AACCGGAGGACAAAGTACA	CAACCGGTGGCGCATTGCC
NCA	TTCCAGAACCGGAGGACAA	AGGAGGAAGGGCCTGAGTC
NCT	GTGTGGTTCCAGAACCGGA	AGCTGGAGGAGGAAGGGCC
NCC	TCCAGAACCGGAGGACAAA	GCATTGCCACGAAGCAGGC
NCG	CAGAAGCTGGAGGAGGAAG	ATTGCCACGAAGCAGGCCA
NGA	CATCAACCGGTGGCGCATT	AGAACCGGAGGACAAAGTA
NGT	GCAGAAGCTGGAGGAGGAA	TCAACCGGTGGCGCATTGC
NGC	CCTCCCTCCCTGGCCCAGG	GAAGCTGGAGGAGGAAGGG

Ejemplo 3: Diversidad de Cas9 y ARN, PAM, dianas

El sistema CRISPR-Cas es un mecanismo inmunitario adaptativo contra la invasión de ADN exógeno empleado por diversas especies a través de bacterias y arqueas. El sistema CRISPR-Cas9 tipo II consiste en un conjunto de genes que codifican proteínas responsables de la "adquisición" de ADN extraño en el locus CRISPR, así como un conjunto de genes que codifican la "ejecución" del mecanismo de escisión del ADN; estos incluyen la nucleasa de ADN (Cas9), un ARN-cr de transactivación no codificante (ARNtracr) y una serie de espaciadores extraídos procedentes de ADN flanqueados por repeticiones directas (ARNcr). Tras la maduración con Cas9, el dúplex ARNtracr y ARNcr guía la nucleasa Cas9 a una secuencia de ADN diana especificada por las secuencias guía espaciadoras, y media rupturas de doble cadena en el ADN cerca de un motivo de secuencia corta en el ADN diana que se requiere para la escisión y específico para cada sistema CRISPR-Cas. Los sistemas CRISPR-Cas tipo II se encuentran en todo el reino bacteriano (Figuras 7 y 8A-F) y son muy diversos en la secuencia y el tamaño de la proteína Cas9, la secuencia de repetición directa de ARNtracr y ARNcr, la organización del genoma de estos elementos y el requisito del motivo para escisión de la diana. Una especie puede tener múltiples sistemas CRISPR-Cas distintos.

Los solicitantes evaluaron 207 Cas9s posibles de especies bacterianas (Figura 8A-F) identificadas basándose en la homología de secuencia con Cas9s conocidos y estructuras ortólogas a subdominios conocidos. Usando el método del Ejemplo 1, los solicitantes llevarán a cabo una evaluación exhaustiva de cada posible desapareamiento en cada posición del ARN guía para que estas Cas9 diferentes generen un modelo para informar el diseño de los ARN guía que tengan una alta especificidad de escisión para cada uno en función del impacto de la posición de prueba y el número de desapareamientos en el ARN guía en la eficacia de escisión para cada Cas9.

El sistema CRISPR-Cas es susceptible de lograr la eliminación dirigida de genes candidatos a enfermedades específica del tejido y controlada temporalmente. Los ejemplos incluyen, pero no se limitan a, genes involucrados en el metabolismo del colesterol y de los ácidos grasos, enfermedades amiloides, enfermedades negativas dominantes, infecciones virales latentes, entre otros trastornos. Por consiguiente, las secuencias diana pueden estar en genes de enfermedad candidatos, por ejemplo:

Enfermedad	GEN	ESPACIADOR	PAM	Mecanismo	Referencias
Hipercolesterolemia	HMG-CR	GCCAAATTGGACGACCCTCG	CGG	Desactivado	Fluvastatin: a review of its pharmacology and use in the management of hypereholesterolaemia.(Plosker GL et al. Drugs 1996, 51(3):433-459)
Hipercolesterolemia	SQLE	CGAGGAGACCCCCGTTTCGG	TGG	Desactivado	Potential role of nonstatin cholesterol lowering agents (Trapani et al. IUBMB Life, Volume 63, Tema 11, páginas 964-971, noviembre de 2011)

Hiperlipidemia	DGAT 1	CCCGCCGCC GCCGTGGCT CG	AGG	Desactivado	DGAT 1 inhibitors as anti-obesity and anti-diabetic agents. (Birch AM et al. Current Opinion in Drug Discovery & Development [2010, 13(4):489-496)
Leucemia	BCR-ABL	TGAGCTCTA CGAGATCCA CA	AGG	Desactivado	Killing of leukemic cells with a BCR/ABL fusion gene by RNA interference (RNAi).(Fuchs et al. Oncogene 2002, 21(37):5716-5724)

Ejemplos de un par de ARN guía para introducir la microdeleción cromosómica en un locus genético

Enfermedad	**GEN**	**ESPACIADOR**	**PAM**	**Mecanismo**	**Referencias**
Hiperlipidemia	guía PLIN2 1	CTCAAAATT CATACCGGT TG	TGG	Microdeleción	Perilipin-2 Null Mice are Protected Against Diet-Induced Obesity, Adipose Inflammation and Fatty Liver Disease (McManaman JL et al. The Journal of Lipid Research, jlr.M035063. First Published on February Id, 2013)
Hiperlipidemia	guía PLIN2	CGTTAAACA ACAACCGGA CT	IGG	Microdeleción	
Hiperlipidemia	guía SRJEBP 1	TTCACCCCG CGGCGCTGA AT	ggg	Microdeleción	Inhibition of SREBP by a Small Molecule, Betulin, Improves Hyperlipidemia and Insulin Resistance and Reduces Atherosclerotic Plaques (Tang I et al. Cell Metabolism, Volume 13, Issue 1, 44-56, 5 January 2011)
Hiperlipidemia	guía SREBP 2	ACCACTACC AGTCCGTCC AC	agg	Microdeleción	

Ejemplos de posibles espaciadores dirigidos contra el VIH-1 incluidos en Mcintyre *et al.,* que generaron ARNsh contra el VIH-1 optimizados para una cobertura máxima de las variantes del VIH-1.
CACTGCTTAAGCCTCGCTCGAGG TCACCAGCAATATTCGCTCGAGG CACCAGCAATATTCCGCTCGAGG TAGCAACAGACATACGCTCGAGG GGGCAGTAGTAATACGCTCGAGG CCAATTCCCATACATTATTGTAC

Identificación del sitio diana Cas9; los solicitantes analizaron el locus genómico CFTR humano e identificaron el sitio diana Cas9 (PAM puede contener un motivo NGG o NNAGAAW). La frecuencia de estas secuencias PAM en el genoma humano se muestra en la figura 5.

Los ID del protoespaciador y su correspondiente diana genómica, la secuencia del protoespaciador, la secuencia PAM y ubicación de la cadena se proporcionan en la siguiente tabla. Las secuencias guía se diseñaron para ser complementarias a la secuencia del protoespaciador completa en el caso de transcritos separados en el sistema híbrido, o solo a la parte subrayada en el caso de los ARN quiméricos.

Tabla: Los ID del protoespaciador y su correspondiente diana genómica, la secuencia del protoespaciador, la secuencia PAM y ubicación de la cadena

ID del protoespaciador	**diana genómica**	**secuencia del protoespaciador (5' a 3')**	**PAM**	**cadena**
1	*EMX1*	GGACATCGATGTCACCTCCAATGACTAG GG	TGG	+

2	*EMX1*	CATTGGAGGTGACATCGATGTCCTCCCC AT	TGG	-
3	*EMX1*	GGAAGGGCCTGAGTCCGAGCAGAAGAA GAA	GGG	+
4	*PVALB*	GGTGGCGAGAGGGGCCGAGATTGGGTGT TC	AGG	+
5	*PVALB*	ATGCAGGAGGGTGGCGAGAGGGGCCGA GAT	TGG	+

Identificación computacional de sitios diana CRISPR únicos: Para identificar sitios diana únicos para una Cas, por ejemplo, una Cas9, por ejemplo, la enzima SF370 Cas9 de *S. pyogenes* (SpCas9), en moléculas de ácido nucleico, por ejemplo, de células, por ejemplo, de organismos, que incluyen, pero no se limitan a, seres humanos, ratón, rata, pez cebra, mosca de la fruta y genoma de *C. elegans*, los solicitantes desarrollaron un paquete de software para escanear ambas cadenas de una secuencia de ADN e identificar todos los posibles sitios diana de SpCas9. En cuanto a este ejemplo, cada sitio diana de SpCas9 se definió operativamente como una secuencia de 20 pb seguida de una secuencia de motivo adyacente al protoespaciador (PAM) NGG, y se identificaron todas las secuencias que satisfacen esta definición 5'-N_{20}-NGG-3' en todos los cromosomas. Para evitar la edición no específica del genoma, después de identificar todos los sitios potenciales, todos los sitios diana se filtraron en función del número de veces que aparecen en el genoma de referencia relevante. Para aprovechar la especificidad de secuencia de Cas, por ejemplo, la actividad de Cas9 conferida por una secuencia 'semilla', que pueden ser, por ejemplo, aproximadamente la secuencia de 11-12 pb en 5' de la secuencia PAM, se seleccionaron las secuencias 5'-NNNNNNNNNN-NGG-3' para que sean únicas en el genoma relevante. Las secuencias genómicas están disponibles en UCSC Genome Browser y en las visualizaciones de muestra de la información del genoma humano hg, genoma de ratón mm, genoma de rata rn, genoma de pez cebra danRer, genoma de *D. melanogaster* dm, genoma de *C. elegans* ce, el genoma porcino y el genoma de vaca se muestran en las figura 15 a 22 respectivamente.

Se puede llevar a cabo un análisis similar para otras enzimas Cas utilizando sus respectivas secuencias PAM, por ejemplo, Cas9 de Staphylococcus aureus sp. Aureus y su secuencia PAM NNGRR (Fig. 31).

Ejemplo 4: Arquitectura experimental para evaluar la actividad diana y la especificidad de CRISPR-Cas

Las nucleasas dirigidas, como los sistemas CRISPR-Cas para aplicaciones de edición de genes, permiten una modificación altamente precisa del genoma. Sin embargo, la especificidad de las herramientas de edición de genes es una consideración crucial para evitar la actividad adversa inespecífica. Aquí, los solicitantes describen un algoritmo de selección de ARN guía Cas9 que predice sitios inespecíficos para cualquier sitio diana deseado dentro de los genomas de mamíferos.

Los solicitantes construyeron grandes bibliotecas de oligos de ARN guía que llevan combinaciones de mutaciones para estudiar la dependencia de secuencia de la programación de Cas9. Utilizando la secuenciación profunda de próxima generación, los solicitantes estudiaron la capacidad de mutaciones únicas y múltiples combinaciones de desapareamientos dentro de diferentes ARN guía de Cas9 para mediar en la modificación del locus del ADN diana. Los solicitantes evaluaron los sitios candidatos inespecíficos con homología de secuencia con el sitio diana de interés para evaluar cualquier escisión inespecífica.

Algoritmo para predecir la actividad diana y la especificidad de CRISPR-Cas: Los datos de estos estudios se utilizaron para desarrollar algoritmos para la predicción de la actividad inespecífica de CRISPR-Cas a través del genoma humano. La plataforma computacional resultante de los solicitantes admite la predicción de toda la actividad diana y la especificidad del sistema CRISPR-Cas en cualquier genoma. Los solicitantes evalúan la actividad y la especificidad de CRISPR-Cas prediciendo la eficacia de corte de Cas9 para cualquier direccionamiento de CRISPR-Cas frente a todas las demás dianas genómicas de CRISPR-Cas, excluyendo los factores limitantes, es decir, algunas modificaciones epigenéticas como la cromatina represiva/heterocromatina.

Los algoritmos que describen los solicitantes 1) evalúan cualquier sitio diana y dan posibles dianas inespecíficas y 2) generan sitios diana candidatos para cualquier locus de interés con una actividad inespecífica mínima prevista.

Modelo termodinámico de cuadrados mínimos de eficacia de corte de CRISPR-Cas: Para sitios diana de Cas9 arbitrarios, los solicitantes generaron un modelo termodinámico numérico que predice la eficacia de corte de Cas9. Los solicitantes proponen 1) que el ARN guía de Cas9 tiene energías libres específicas de hibridación a su diana y cualquier secuencia de ADN inespecífica y 2) que Cas9 modifica las energías libres de hibridación de ARN:ADN localmente de una manera dependiente de la posición pero independiente de la secuencia. Los solicitantes captaron un modelo para predecir la eficacia de corte de CRISPR-Cas basándose en sus datos de mutación de ARN guía de CRISPR-Cas y en los cálculos de energía termodinámica ARN:ADN utilizando un algoritmo de aprendizaje automático.

Luego, los solicitantes validaron sus modelos resultantes mediante la comparación de sus predicciones de corte inespecíficos de CRISPR-Cas en múltiples loci genómicos con datos experimentales que evalúan la modificación de locus en los mismos sitios.

La metodología adoptada en el desarrollo de este algoritmo es la siguiente: El resumen del problema indica que para espaciadores arbitrarios y dianas de longitud constante, se debe encontrar un modelo numérico que tenga sentido termodinámico y prediga la eficacia de corte de Cas9.

Suponiendo que Cas9 modifica las energías libres de hibridación de ADN:ARN localmente de forma dependiente de la posición pero independiente de la secuencia. Luego, para las energías libres de hibridación de ADN:ARN $\Delta G_{ij}(k)$ (para la posición k entre 1 y N) del espaciador i y de la diana j

$$Z_{ij} = \sum_{k=1}^{N} \alpha_k \Delta G_{ij}(k)$$

Z_{ij} se trata como una energía libre "eficaz" modificada por la posición multiplicativa-pesos α_k.

El Z_{ij} "eficaz" de energía libre corresponde a una probabilidad de corte asociada ~ $e^{-\beta Z_{ij}}$ (para algunos constantes β) de la misma manera que un modelo de equilibrio de hibridación (sin ponderación de posición) habría predicho una probabilidad de hibridación ~$e^{-\beta \Delta G_{ij}}$. Dado que se ha medido la eficacia de corte, los valores Z_{ij} pueden tratarse como sus observables. Paralelamente, se puede calcular $\Delta G_{ij}(k)$ para el emparejamiento espaciador-diana de cualquier experimento. La tarea de los solicitantes era encontrar los valores α_k, ya que esto les permitiría estimar Z_{ij} para cualquier par espaciador-diana.

Escribiendo la ecuación anterior para Z_{ij} en forma matricial, los solicitantes obtienen:

$$\vec{Z} = \mathbf{G}\vec{\alpha} \qquad (1)$$

La estimación de mínimos cuadrados es entonces

$$\vec{\alpha}_{est} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\vec{Z}$$

donde G^T es la matriz-transposición de la G y $(G^TG)^{-1}$ es la inversa de su matriz-producto.

En la anterior, **G** es una matriz de valores locales de energía libre de ADN:ARN cuya fila r corresponde al ensayo experimental r y cuya columna k corresponde a la posición k en el híbrido de ADN:ARN probado en ese ensayo experimental. $\vec{Z}$ es, mientras tanto, un vector-columna cuya fila r corresponde a los observables del mismo ensayo experimental que la fila r de **G**. Debido a la relación descrita anteriormente en la que se estima que las frecuencias de corte de CRISPR varían según ~ $e^{-\beta Z_{ij}}$, estos observables, Z_{ij}, se calcularon como el logaritmo natural de la frecuencia de corte observada. Lo observable es la eficacia de escisión de Cas, por ejemplo, Cas9, en un ADN diana para un ARN guía particular y un par de ADN dianas. El experimento es Cas, por ejemplo, Cas9, con un emparejamiento ARNsg/ADN diana particular, y el observable es el porcentaje de escisión (ya sea medido como porcentaje de formación de indel de células o simplemente porcentaje de escisión *in vitro*) (véase en el presente documento la discusión sobre cómo generar un conjunto de datos de capacitación). Más en particular, cada reacción de PCR única que fue secuenciada debe tratarse como un ensayo experimental único para abarcar la replicabilidad dentro del vector. Esto significa que cada réplica experimental va en filas separadas de la ecuación 1 (y debido a esto, algunas filas de **G** serán idénticas). La ventaja de esto es que cuando $\vec{\alpha}$ se ajusta, toda la información relevante, incluida la replicabilidad, se tiene en cuenta en la estimación final.

Los valores $\vec{Z}$ observables, se calcularon como log (frecuencia de corte observada). Las frecuencias de corte se normalizaron de manera idéntica (de modo que todas tienen las mismas "unidades"). Sin embargo, para conectar valores de frecuencia indel de secuencia, puede ser mejor estandarizar la profundidad de secuenciación.

La forma preferida de hacer esto sería establecer una profundidad de secuencia estándar D para la cual todos los experimentos incluidos en $\vec{Z}$ tengan al menos ese número de lecturas. Debido a que las frecuencias de corte por debajo de 1/D no se pueden detectar de manera consistente, debe establecerse como la frecuencia mínima para el conjunto de datos, y los valores en $\vec{Z}$ deben variar desde log(1/D) a log(1). Se podría variar el valor de D más adelante para asegurarse de que la estimación de $\vec{\alpha}$ no sea demasiado dependiente del valor elegido.
En un aspecto adicional, existen diferentes métodos para graficar secuencias NGG y NNAGAAW. Uno es con el

método 'no superpuesto'. NGG y NRG pueden registrarse de forma "superpuesta", como se indica en las figuras 6A-C.

Los solicitantes también realizaron un estudio sobre la actividad de Cas9 inespecífica como se indica en las figuras 10, 11 y 12. Los aspectos de la invención también se relacionan con modelos predictivos que pueden no involucrar energías de hibridación, sino que simplemente usan la información de frecuencia de corte como una predicción (Véase la figura 29).

Ejemplo 5: ADN. Especificidad dirigida de la nucleasa Cas9 guiada por ARN

Aquí, los solicitantes informan sobre la optimización de varias aplicaciones de SpCas9 para la edición del genoma de mamíferos y demuestran que la escisión mediada por SpCas9 no se ve afectada por la metilación del ADN (Fig. 14). Los solicitantes caracterizan además la especificidad dirigida de SpCas9 utilizando más de 700 variantes de ARN guía y evalúan los niveles de mutación indel inducidos por SpCas9 en más de 100 loci genómicos inespecíficos predichos. Contrariamente a los modelos anteriores, los solicitantes encontraron que SpCas9 tolera los desapareamientos entre el ARN guía y el ADN diana en diferentes posiciones de una manera dependiente del contexto de la secuencia, sensible al número, la posición y la distribución de los desapareamientos. Por último, los solicitantes demuestran que la dosis de SpCas9 y ARNsg se puede ajustar para minimizar la modificación inespecífica. Para facilitar las aplicaciones de ingeniería del genoma de los mamíferos, los solicitantes utilizaron estos resultados para establecer una plataforma computacional para guiar la selección y validación de las secuencias diana, así como los análisis inespecíficos.

El sistema bacteriano CRISPR tipo II de *S. pyogenes* puede reconstituirse en células de mamíferos utilizando tres componentes mínimos: la nucleasa Cas9 (SpCas9), un ARN CRISPR determinante de la especificidad (ARNcr), y un ARNcr transactivador auxiliar (ARNtracr). Después de la hibridación de ARNcr y ARNtracr, SpCas9 se localiza en la diana genómica que coincide con una secuencia guía de 20 nt dentro del ARNcr, inmediatamente cadena arriba de un motivo adyacente al protoespaciador (PAM) 5'-NGG requerido. Cada dúplex ARNcr y ARNtracr también se pueden fusionar para generar un ARN guía quimérico único. (ARNsg.) que imita el híbrido ARNcr-ARNtracr natural. Tanto los dúplex ARNcr-ARNtracr como los ARNsg se pueden usar para dirigir a SpCas9 para la edición del genoma multiplexado en células eucariotas.

Aunque se había demostrado previamente que un diseño de ARNsg, que consiste en un ARNcr truncado y ARNtracr, medía una escisión eficaz *in vitro*, falló para lograr una escisión detectable en varios loci que fueron modificados de manera eficaz mediante dúplex ARNcr-ARNtracr que llevan secuencias guía idénticas. Debido a que la principal diferencia entre este diseño de ARNsg y el dúplex ARNcr-ARNtracr nativo es la longitud de la secuencia del ARNtracr, los solicitantes probaron si la extensión de la cola del ARNtracr era capaz de mejorar la actividad de SpCas9.

Los solicitantes generaron un conjunto de ARNsg dirigidos a múltiples sitios dentro de los loci humanos EMX1 y PVALB con diferentes truncamientos en 3' de ARNtracr. Usando el ensayo de nucleasa SURVEYOR, los solicitantes evaluaron la capacidad de cada complejo de ARNsg de Cas9 para generar indel en células HEK 293 FT a través de la inducción de roturas bicatenarias de ADN (DSB) y la subsiguiente reparación de daños en el ADN mediante uniones terminales no homólogas (NHEJ) (Métodos y Materiales). Los ARNsg con colas de ARNtracr de +67 o +85 nucleótidos (nt) mediadas por la escisión del ADN en todos los sitios diana analizados, con niveles hasta 5 veces más altos de indeles que los dúplex de ARNcr-ARNtracr correspondientes. Además, ambos diseños de ARNsg modificaron de manera eficaz los loci de PVALB que anteriormente no se podían utilizar de forma dirigida con los dúplex de ARNcr-ARNtracr. Para las cinco dianas probadas, los solicitantes observaron un aumento constante en la eficacia de modificación al aumentar la longitud del ARNtracr. Los solicitantes realizaron transferencias Northern para los truncamientos del ARN guía y encontraron niveles de expresión aumentados para las secuencias más largas del ARNtracr, lo que sugiere que la escisión de la diana mejorada se debió a una mayor expresión o estabilidad del ARNsg. En conjunto, estos datos indican que la cola del ARNtracr es importante, para la expresión y actividad óptima de SpCas9 *in vivo*.

Los solicitantes investigaron adicionalmente la arquitectura del ARNsg extendiendo la longitud del dúplex desde 12 hasta los 22 nt encontrados en el dúplex ARNcr-ARNtracr nativo. Los solicitantes también mutaron la secuencia que codifica el ARNsg para suprimir cualquier vía poli-T que podría servir como terminadores de la transcripción prematuros para la transcripción dirigida por U6. Los solicitantes probaron estos nuevos armazones de ARNsg en 3 dianas dentro del gen EMX1 humano y solo observaron cambios modestos en la eficacia de la modificación. Por lo tanto, los solicitantes establecieron que el ARNsg (+85), idéntico a algunos ARNsg previamente utilizados, como una arquitectura de ARN guía de SpCas9 eficaz y lo usó en todos los estudios posteriores.

Los solicitantes han demostrado previamente que un mutante catalítico de SpCas9 (nickasa D10A) puede mediar la edición de genes mediante reparación dirigida por homología (HR) sin formación de indel detectable. Dada su mayor eficacia de escisión, los solicitantes probaron si el ARNsg (+85), en complejo con la nickasa Cas9, también puede facilitar la HR sin incurrir en NHEJ en la diana. Utilizando oligonucleótidos monocatenarios (ODNss) como plantillas de reparación, los solicitantes observaron que tanto el tipo silvestre como el SpCas9 D10A median la HR en las células HEK 293FT, mientras que solo el primero es capaz de hacerlo en células madre embrionarias humanas. Los solicitantes confirmaron además, utilizando el ensayo SURVEYOR, que no se induce ninguna mutación indel diana

por la nickasa SpCas9 D10A.

Para explorar si la capacidad genómica dirigida del ARNsg (+85) está influenciada por factores epigenéticos que limitan las tecnologías de nucleasas efectoras tipo activador de transcripción (TALEN) alternativas y potencialmente también de nucleasas de dedos de zinc (ZFN). Los solicitantes probaron además la capacidad de SpCas9 para escindir el ADN metilado. Utilizando pUC19 no metilado o M. Sssl metilado como dianas de ADN (Fig. 14a, b) en un ensayo de escisión libre de células, los solicitantes demostraron que SpCas9 escinde eficazmente pUC19 independientemente del estado de metilación de CpG en la secuencia diana de 20 pb o el PAM (Fig. 14c). Para probar si esto también es cierto *in vivo*, los solicitantes diseñaron ARNsg para dirigir a una región altamente metilada del locus SERPINB5 humano. Los tres ARNsg probados fueron capaces de mediar mutaciones indel en dianas metiladas endógenamente.

Una vez establecida la arquitectura de ARN guía óptima para SpCas9 y demostrada su insensibilidad a la metilación de CpG genómica, los solicitantes intentaron realizar una caracterización completa de la especificidad de ADN dirigida a SpCas9. Los estudios previos sobre la especificidad de escisión de SpCas9 se limitaron a un pequeño conjunto de desapareamientos de un solo nucleótido entre la secuencia guía y la diana de ADN, lo que sugiere que el emparejamiento de bases perfecto dentro de 10-12 pb directamente en el 5' de PAM determina la especificidad de Cas9, mientras que se pueden tolerar los desapareamientos múltiples PAM-distal. Además, un estudio reciente que utiliza SpCas9 catalíticamente inactivo como represor transcripcional no encontró efectos significativos inespecíficos en todo el transcriptoma de *E. coli*. Sin embargo, aún no se ha informado un análisis sistemático de la especificidad de Cas9 en el contexto de un genoma de mamífero más grande.

Para abordar esto, los solicitantes primero evaluaron el efecto de la identidad de ARN guía imperfecta para dirigirse al ADN genómico en la actividad de SpCas9, y luego evaluaron la actividad de escisión resultante de un único ARNsg en múltiples loci genómicos inespecíficos con similitud de secuencia. Para facilitar las pruebas a gran escala de secuencias guía desparejadas, los solicitantes desarrollaron un sencillo ensayo de prueba de ARNsg mediante la generación de casetes de expresión que codifican ARNsg dirigidos por U6 mediante PCR y la transfección de los amplicones resultantes. Los solicitantes luego realizaron una secuenciación profunda de la región que flanqueaba cada sitio diana para dos réplicas biológicas independientes. A partir de estos datos, los solicitantes aplicaron un modelo binomial para detectar verdaderos eventos de indel resultantes de la escisión de SpCas9 y la mala reparación del NHEJ y calcularon intervalos de confianza del 95 % para todas las frecuencias NHEJ indicadas.

Los solicitantes utilizaron un modelo lineal de dependencia de la posición de energía libre para investigar la contribución combinada de la secuencia de ADN:ARN y la ubicación de desapareamientos en la eficacia de corte de Cas9. Si bien la composición de la secuencia y la ubicación de los desapareamientos solo generaron correlaciones de Spearman entre las eficacias de corte estimadas y observadas para el sitio diana I EMX1 y .78, respectivamente, la integración de los dos parámetros mejoró enormemente este acuerdo, con la correlación de Spearman .86 ($p < 0,001$). Además, la incorporación de las energías de hibridación de ARN nupac:ARN en el modelo de energía libre de los solicitantes dio como resultado un aumento del 10 % en el coeficiente de correlación de Spearman. En conjunto, los datos sugieren un efecto de las perturbaciones específicas de SpCas9 en las energías libres de emparejamiento de bases de Watson-Crick. Paralelamente, la composición de la secuencia no mejoró sustancialmente la concordancia entre las eficacias de corte estimadas y observadas para el sitio diana 6 de EMX1 (correlación de Spearman 0.91, $p < 0,001$). Esto sugirió que los desapareamientos únicos en el sitio diana 6 de EMX1 contribuyeron mínimamente a la energía libre de unión termodinámica en sí.

Los sitios genómicos inespecíficos potenciales con una similitud de secuencia a un sitio diana de interés a menudo pueden tener múltiples desapareamientos de base. Los solicitantes diseñaron un conjunto de ARN guía para las dianas 1 y 6 de EMX1 que contienen diferentes combinaciones de desapareamientos para investigar el efecto del número, la posición y la separación de desapareamientos en la actividad de escisión de la diana Cas9 (Fig. 13a, b).

Al concatenar bloques de desapareamientos, los solicitantes encontraron que dos desapareamientos consecutivos dentro de la secuencia proximal de PAM redujeron el corte de Cas9 para ambas diana a < 1 % (Fig. 13a; paneles superiores). El corte en el sitio diana 1 aumentó a medida que los desapareamientos dobles se desplazaron distalmente desde PAM, mientras que la escisión observada para el sitio diana 6 permaneció consistentemente < 0,5 %. Los bloques de tres o cinco desapareamientos consecutivos para ambas dianas disminuyeron el corte de Cas9 a niveles < 0,5 % independientemente de la posición (Fig. 13, paneles inferiores).

Para investigar el efecto de la separación de desapareamientos, los solicitantes anclaron una única mutación proximal de PAM mientras aumentaban sistemáticamente la separación entre desapareamientos posteriores. Grupos de 3 o 4 mutaciones separadas cada una por 3 o menos bases disminuyeron la actividad de nucleasa Cas9 a niveles < 0,5 %. Sin embargo, el corte de Cas9 en el sitio diana 1 aumentó a 3-4 % cuando las mutaciones se separaron por 4 o más bases no mutadas (Fig. 13b). De manera similar, grupos de 4 mutaciones separadas por 4 o más bases condujeron a eficacias de indel del 0,5-1 %. Sin embargo, la escisión en el sitio diana 6 permaneció consistentemente por debajo del 0,5 %, independientemente del número o la separación de los desapareamientos del ARN guía.

La datos múltiples de desapareamientos del ARN guía indican que el aumento del número de mutaciones disminuye y eventualmente elimina la escisión. Inesperadamente, se toleran mutaciones aisladas a medida que aumenta la

separación entre cada desapareamiento. De acuerdo con los datos de un solo desapareamiento, las mutaciones múltiples dentro de la región distal de PAM generalmente son toleradas por Cas9 mientras que los grupos de mutaciones proximales de PAM no lo son. Por último, aunque las combinaciones de desapareamientos representan un subconjunto limitado de mutaciones de base, parece haber una susceptibilidad específica a la diana para guiar los desapareamientos de ARN. Por ejemplo, el sitio diana 6 generalmente mostró una escisión más baja con múltiples desapareamientos, una propiedad que también se refleja en su región proximal PAM más larga de 12-14 pb de intolerancia a la mutación (Fig. 12). Una investigación adicional de la especificidad de la secuencia de Cas9 puede revelar pautas de diseño para elegir dianas de ADN más específicas.

Para determinar si los hallazgos de los solicitantes a partir de los datos de la mutación del ARN guía se generalizan para identificar los desapareamientos del ADN diana y permiten la predicción de la escisión inespecífica dentro del genoma, los solicitantes transfectaron células con Cas9 y ARN guía dirigidos a la diana 3 o a la diana 6, y realizaron una secuenciación profunda de sitios candidatos inespecíficos con similitud de secuencia. No se identificaron loci genómicos con solo 1 desapareamiento con ninguna de las dianas. Los loci genómicos que contenían 2 o 3 desapareamientos en relación con la diana 3 o la diana 6 revelaron escisión en algunos de las dianas evaluadas (Fig. 13c). Las dianas 3 y 6 mostraron eficacias de escisión del 7,5 % y del 8,0 %, mientras que los sitios inespecíficos 3-1, 3-2, 3-4 y 3-5 se modificaron al 0,19 %, 0,42 %, 0,97 % y 0,50 %, respectivamente. Todos los demás sitios inespecíficos se escindieron en menos del 0,1 % o se modificaron en niveles que no se distinguen del error de secuenciación. Las tasas de corte inespecíficas fueron consistentes con los resultados colectivos de los datos de mutación del ARN guía: la escisión se observó en un pequeño subconjunto de la diana 3 inespecífica que contenía desapareamientos muy distales de PAM o tenían desapareamientos únicos separados por 4 o más bases.

Dado que las eficacias dirigidas al genoma de las TALEN y ZFN pueden ser sensibles a los efectos de confusión, como el estado de la cromatina o la metilación del ADN, los solicitantes intentaron probar si la actividad de escisión SpCas9 guiada por ARN se vería afectada por el estado epigenético de un locus diana. Para ensayar esto, los solicitantes metilaron un plásmido *in vitro* y realizaron un ensayo de escisión *in vitro* en dos pares de dianas que contenían CpG no metilados o metilados. SpCas9 medió la escisión eficaz del plásmido si la metilación ocurrió en la diana propiamente dicha o dentro del PAM, lo que sugiere que SpCas9 puede no ser susceptible a los efectos de metilación del ADN.

La capacidad de programar Cas9 para dirigirse a sitios específicos en el genoma simplemente mediante el diseño de un ARNsg corto tiene un enorme potencial para varias aplicaciones. Los resultados de los solicitantes demuestran que la especificidad de la escisión del ADN mediada por Cas9 depende de la secuencia y se rige no solo por la ubicación de las bases desapareadas, sino también por su separación. De manera importante, mientras que el PAM de 9-12 nt proximal de la secuencia guía generalmente define la especificidad, las secuencias PAM distales también contribuyen a la especificidad general de la escisión del ADN mediada por Cas9. Aunque hay sitios de escisión inespecíficos para una secuencia de guía dada, es probable que los sitios inespecíficos esperados sean predecibles en función de sus ubicaciones de desapareamientos. El trabajo adicional que analiza la termodinámica de la interacción de ARNsg-ADN probablemente proporcionará un poder predictivo adicional para la actividad inespecífica, y la exploración de ortólogos de Cas9 alternativos también puede producir nuevas variantes de Cas9s con especificidad mejorada. En conjunto, la alta eficacia de Cas9 así como su baja actividad inespecífica hacen de CRISPR-Cas una atractiva tecnología de ingeniería genómica.

Ejemplo 6: Uso de Cas9 para dirigir varios tipos de enfermedades

La especificidad de los ortólogos de Cas9 se puede evaluar probando la capacidad de cada Cas9 para tolerar desapareamientos entre el ARN guía y su diana de ADN. Por ejemplo, la especificidad de SpCas9 se ha caracterizado probando el efecto de las mutaciones en el ARN guía sobre la eficacia de la escisión. Se crearon bibliotecas de ARN guía con desapareamientos únicos o múltiples entre la secuencia guía y el ADN diana. Basándose en estos hallazgos, los sitios diana para SpCas9 se pueden seleccionar según las siguientes pautas:

Para maximizar la especificidad de SpCas9 para editar un gen en particular, se debe elegir un sitio diana dentro del locus de interés, de modo que las posibles secuencias genómicas inespecíficas cumplan con las siguientes cuatro restricciones: En primer lugar y principalmente, no deben ir seguidos de un PAM con secuencias 5'-NGG o NAG. En segundo lugar, su secuencia global similar a la diana, la secuencia debe minimizarse. En tercer lugar, un número máximo de desapareamientos debe estar dentro de la región proximal de PAM del sitio inespecífico. Por último, un número máximo de desapareamientos debe ser consecutivo o espaciado a menos de cuatro bases.

Se pueden usar métodos similares para evaluar la especificidad de otros ortólogos de Cas9 y establecer criterios para la selección de sitios diana específicos dentro de los genomas de las especies diana.

Selección de dianas para ARNsg: Hay dos consideraciones principales en la selección de la secuencia guía de 20 nt para la orientación de genes: 1) la secuencia diana debe preceder al PAM 5'-NGG para Cas9 de S. pyogenes, y 2) las secuencias guía deben elegirse para minimizar la actividad inespecífica. Los solicitantes proporcionaron una herramienta de diseño de orientación Cas9 en línea (disponible en el sitio web genome-engineering.org/tools; véanse los ejemplos anteriores y la figura 23) que toma una secuencia de entrada de interés e identifica los sitios diana

adecuados. Para evaluar experimentalmente las modificaciones inespecíficas para cada ARNsg, los solicitantes también proporcionan sitios inespecíficos pronosticados computacionalmente para cada diana prevista, clasificados de acuerdo con el análisis de especificidad cuantitativa de los solicitantes sobre los efectos de la identidad, posición y distribución de desapareamiento de emparejamientos de bases.

La información detallada sobre los sitios inespecíficos pronosticados computacionalmente es la siguiente: Consideraciones para las actividades de escisión inespecífica: Al igual que otras nucleasas, Cas9 puede escindir dianas de ADN inespecíficas en el genoma a frecuencias reducidas. El grado en que una secuencia guía dada exhibe actividad inespecífica depende de una combinación de factores que incluyen la concentración de enzima, las termodinámicas de la secuencia guía específica empleada y la abundancia de secuencias similares en el genoma diana. Para la aplicación rutinaria de Cas9, es importante considerar formas de minimizar el grado de escisión inespecífica y también ser capaz de detectar la presencia de escisión inespecífica.

Minimizar la actividad inespecífica: Para la aplicación en líneas celulares, los solicitantes recomiendan seguir dos pasos para reducir el grado de modificación del genoma inespecífica. En primer lugar, al utilizar la herramienta de selección de direccionamiento de CRISPR en línea de los solicitantes, es posible evaluar computacionalmente la probabilidad de que una secuencia guía dada tenga sitios inespecíficos. Estos análisis se realizan mediante una búsqueda exhaustiva en el genoma de secuencias inespecíficas que son secuencias similares a la secuencia guía. La investigación experimental exhaustiva del efecto de las bases desapareadas entre el ARNsg y su ADN diana reveló que la tolerancia de desapareamientos es 1) dependiente de la posición - las 8-14 pb en el extremo 3' de la secuencia guía son menos tolerantes a los desapareamientos que las bases en el 5', 2) dependiente de la cantidad - en general, no se toleran más de 3 desapareamientos, 3) dependiente de la secuencia guía - algunas secuencias de guía son menos tolerantes a los desapareamientos que otras, y 4) dependiente de la concentración - la escisión inespecífica es altamente sensible a la cantidad de ADN transfectado. La herramienta web de análisis del sitio diana de los solicitantes (disponible en el sitio web genome-engineering.org/tools) integra estos criterios para proporcionar predicciones para los posibles sitios inespecíficos en el genoma diana. En segundo lugar, los solicitantes recomiendan valorar la cantidad de plásmido de expresión Cas9 y ARNsg para minimizar la actividad inespecífica.

Detección de actividades inespecíficas: al utilizar la herramienta web de direccionamiento de CRISPR de los solicitantes, es posible generar una lista de los sitios inespecíficos más probables, así como los cebadores que ejecutan SURVEYOR, o el análisis de secuencia de esos sitios. Para los clones isogénicos generados con Cas9, los solicitantes recomiendan encarecidamente la secuenciación de estos sitios inespecíficos para verificar si hay mutaciones no deseadas. Vale la pena señalar que puede haber modificaciones inespecíficas en los sitios que no están incluidos en la lista de candidatos predichos y se debe realizar una secuencia completa del genoma para verificar completamente la ausencia de sitios inespecíficos. Además, en los ensayos multiplex en los que se inducen varios DSB dentro del mismo genoma, puede haber tasas bajas de eventos de translocación y se pueden evaluar utilizando varias técnicas como la secuenciación profunda (48).

La herramienta en línea (Fig. 23) proporciona las secuencias para todos los oligos y cebadores necesarios para 1) preparar las construcciones de ARNsg, 2) analizar la eficacia de la modificación de la diana y 3) evaluar la escisión en los posibles sitios inespecíficos, vale la pena señalarlo porque el promotor U6 ARN polimerasa III utilizado para expresar el ARNsg prefiere un nucleótido de guanina (G) como la primera base de su transcripción, se agrega una G adicional en el 5' del ARNsg donde la secuencia guía de 20 nt no comienza con G (Fig. 24).

Ejemplo 7: Investigaciones de desapareamiento con el par de bases

los solicitantes probaron si la extensión de la cola del ARNtracr era capaz de mejorar la actividad de SpCas9. Los solicitantes generaron un conjunto de ARNsg dirigidos a múltiples sitios dentro de los loci humanos EMX1 y PVALB con diferentes truncamientos en el 3' de ARNtracr (Fig. 9a). Usando el ensayo de nucleasa SURVEYOR, los solicitantes evaluaron la capacidad de cada complejo de ARNsg de Cas9 para generar indel en células HEK 293FT a través de la inducción de roturas bicatenarias de ADN (DSB) y la subsiguiente reparación de daños en el ADN mediante uniones terminales no homólogas (NHEJ) (Métodos y Materiales). Los ARNsg con colas de ARNtracr de +67 o +85 nucleótidos (nt) mediadas por la escisión del ADN en todos los sitios diana analizados, con niveles hasta 5 veces más altos de indeles que los dúplex de ARNcr-ARNtracr correspondientes (Fig. 9). Además, ambos diseños de ARNsg modificaron de manera eficaz los loci de PVALB que anteriormente no se podían utilizar de forma dirigida con los dúplex de ARNcr-ARNtracr (1) (Fig. 9b y Fig. 9b). Para las cinco dianas probadas, los solicitantes observaron un aumento constante en la eficacia de modificación al aumentar la longitud del ARNtracr. Los solicitantes realizaron transferencias Northern para los truncamientos del ARN guía y encontraron niveles de expresión aumentados para las secuencias más largas del ARNtracr, lo que sugiere que la escisión de la diana mejorada se debió a una mayor expresión o estabilidad del ARNsg (Fig. 9c). En conjunto, estos datos indican que la cola del ARNtracr es importante para la expresión y actividad óptima de SpCas9 *in vivo*.
Los solicitantes han demostrado previamente que un mutante catalítico de SpCas9 (nickasa D10A) puede mediar la edición de genes mediante reparación dirigida por homología (HR) sin formación de indel detectable. Dada su mayor eficacia de escisión, los solicitantes probaron si el ARNsg (+85), en complejo con la nickasa Cas9, también puede facilitar la HR sin incurrir en NHEJ en la diana. Utilizando oligonucleótidos monocatenarios (ODNss) como plantillas de reparación, los solicitantes observaron que tanto el tipo silvestre como el SpCas9 D10A median la HR en las células

HEK 293FT, mientras que solo el primero es capaz de hacerlo en células madre embrionarias humanas (hESC; Fig. 9d).

Para explorar si la capacidad genómica dirigida del ARNsg (+85) está influenciada por factores epigenéticos que limitan las tecnologías de nucleasas efectoras tipo activador de transcripción (TALEN) alternativas y potencialmente también de nucleasas de dedos de zinc (ZFN), los solicitantes probaron además la capacidad de SpCas9 para escindir el ADN metilado. Utilizando pUC19 no metilado o M. Sssl metilado como dianas de ADN (Fig. 14a, b) en un ensayo de escisión libre de células, los solicitantes demostraron que SpCas9 escinde eficazmente pUC19 independientemente del estado de metilación de CpG en la secuencia diana de 20 pb o el PAM. Para probar si esto también es cierto *in vivo*, los solicitantes diseñaron ARNsg para dirigir a una región altamente metilada del locus SERPINB5 humano (Fig. 9e, f). Los tres ARNsg probados fueron capaces de mediar mutaciones indel en dianas metiladas endógenamente (Fig. 9g).

Los solicitantes investigaron sistemáticamente el efecto de los desapareamientos de emparejamiento de bases entre las secuencias de ARN guía y el ADN diana en la eficacia de modificación de la diana. Los solicitantes eligieron cuatro sitios diana dentro del gen EMX1 humano y, para cada uno, generaron un conjunto de 57 ARN guía diferentes que contenían todas las posibles sustituciones de nucleótidos individuales en las posiciones 1-19 directamente en el 5' del PAM NAMP requerido (Fig. 25a). La guanina en el 5' en la posición 20 se conserva, dado que el promotor U6 requiere guanina como la primera base de su transcripción. A continuación, se evaluó la actividad de escisión de estos ARN guía "inespecíficos" en el locus genómico en la diana.

De acuerdo con los hallazgos anteriores, SpCas9 tolera los desapareamientos de base única en la región distal de PAM en mayor medida que en la región proximal de PAM. En contraste con un modelo que implica una secuencia semilla proximal de PAM prototípica de 10-12 pb que determina la especificidad de la diana, los solicitantes encontraron que la mayoría de las bases dentro del sitio diana se reconocen específicamente, aunque los desapareamientos se toleran en diferentes posiciones de una manera dependiente del contexto de secuencia. La especificidad de base única generalmente varía de 8 a 12 pb inmediatamente cadena arriba del PAM, lo que indica un límite de especificidad dependiente de la secuencia que varía en longitud (Fig. 25b).

Para investigar adicionalmente las contribuciones de la identidad y posición de bases dentro del ARN guía a la especificidad de SpCas9, los solicitantes generaron conjuntos adicionales de ARN guía desapareados para otros once sitios diana dentro del locus EMX1 (Fig. 28) por un total de más de 400 ARNsg. Estos ARN guía se diseñaron para cubrir todos los 12 posibles desapareamientos de ARN:ADN para cada posición en la secuencia guía con al menos 2X de cobertura para las posiciones 1-10. Los datos agregados de desapareamientos únicos de los solicitantes revelan múltiples excepciones al modelo de secuencia semilla de la especificidad de SpCas9 (Fig. 25c). En general, los desapareamientos dentro de las 8-12 bases proximales de PAM fueron menos tolerados por SpCas9, mientras que los de las regiones distales de PAM tuvieron poco efecto sobre la escisión de SpCas9. Dentro de la región proximal de PAM, el grado de tolerancia varió con la identidad de un desapareamiento particular, con el par de bases rC:dC que exhibe el nivel más alto de interrupción de la escisión de SpCas9 (Fig. 25c).

Además de la especificidad diana, los solicitantes también investigaron el requisito de PAM NGG de SpCas9. Para variar la segunda y tercera posición de PAM, los solicitantes seleccionaron 32 sitios diana dentro del locus EMX1 que abarcan las 16 PAM alternativos posibles con cobertura 2x (Tabla 4). Usando el ensayo SURVEYOR, los solicitantes demostraron que SpCas9 también escinde dianas con PAM NAG, aunque 5 veces menos eficaces que los sitios diana con PAM NGG (Fig. 25d). La tolerancia para un PAM NAG está de acuerdo con estudios bacterianos previos (12) y expande el espacio diana de Cas9 de *S. pyogenes* a cada 4 pb en promedio dentro del genoma humano, sin tener en cuenta factores restrictivos como la estructura secundaria del ARN guía o ciertas modificaciones epigenéticas (Fig. 25e).

Los solicitantes exploraron a continuación el efecto de los desapareamientos de bases múltiples en la actividad diana de SpCas9. Para cuatro dianas dentro del gen EMX1, los solicitantes diseñaron conjuntos de ARN guía que contenían combinaciones variables de desapareamientos para investigar el efecto del número, la posición y la separación de desapareamientos en la actividad de escisión de la diana de SpCas9 (Fig. 26a, b).

En general, los solicitantes observaron que el número total de pares de bases desapareados es un determinante clave para la eficacia de escisión de SpCas9. Dos desapareamientos, particularmente los que ocurren en una región proximal de PAM, redujeron significativamente la actividad de SpCas9 ya sea que estos desapareamientos estén concatenados o interespaciados (Fig. 26a, b); este efecto se amplía aún más para tres desapareamientos concatenados (Fig. 20a). Además, tres o más desapareamientos interespaciados (Fig. 26c) y cinco concatenados (Fig. 26a) eliminaron la escisión detectable de SpCas9 en la gran mayoría de los loci.

La posición de los desapareamientos dentro de la secuencia guía también afectó la actividad de SpCas9: los desapareamientos proximales de PAM son menos tolerados que los homólogos distales de PAM (Fig. 26a), recapitulando las observaciones de los solicitantes a partir de los datos de desapareamiento de pares de bases individuales (Fig. 25c). Este efecto es particularmente notable en las secuencias guía que llevan un pequeño número de desapareamientos totales, ya sean concatenados (Fig. 26a) o interespaciados (Fig. 26b). Además, las secuencias guía con desapareamientos espaciados con cuatro o más bases separadas también mediaron la escisión de SpCas9

en algunos casos (Fig. 26c). Por lo tanto, junto con la identidad de emparejamiento de bases desapareadas, los solicitantes observaron que muchos efectos de escisión inespecífica pueden explicarse mediante una combinación de posición y número de desapareamientos.

Teniendo en cuenta estos resultados de ARN guía desapareados, los solicitantes esperaban que para cualquier ARNsg en particular, SpCas9 pueda escindir loci genómicos que contengan pequeñas cantidades de bases desapareadas. Para las cuatro dianas de EMX1 descritas anteriormente, los solicitantes identificaron computacionalmente 117 sitios inespecíficos en el genoma humano seguidos de un PAM 5'-NRG y cumplen con cualquiera de los siguientes criterios adicionales: 1. hasta 5 desapareamientos, 2. inserciones o deleciones cortas o 3, desapareamientos solo en la región distal de PAM. Además, los solicitantes evaluaron loci inespecíficos de alta similitud de secuencia sin el requisito de PAM. La mayoría de los sitios inespecíficos analizados para cada ARNsg (sitios 30/31, 23/23, 48/51 y 12/12 para las dianas EMX1 1, 2, 3 y 6, respectivamente) mostraron eficacias de modificación al menos 100 veces más bajas que la de las correspondientes dianas (Fig. 27a, b). De los cuatro sitios inespecíficos identificados, tres contenían solo desapareamientos en la región distal de PAM, lo que concuerda con las observaciones de ARNsg de desapareamientos múltiples de los solicitantes (Fig. 26). De forma destacable, estos tres loci fueron seguidos por PAM 5'-NAG, lo que demuestra que los análisis inespecíficos de SpCas9 deben incluir 5'-NAG, así como candidatos loci de 5'-NGG.

La especificidad enzimática y la fuerza de la actividad a menudo dependen en gran medida de las condiciones de reacción, que a una alta concentración de reacción podría amplificar la actividad inespecífica (26, 27). Una estrategia potencial para minimizar la escisión no específica es limitar la concentración de la enzima, a saber, el nivel de complejo SpCas9-ARNsg. La especificidad de la escisión, medida como una relación de escisión específica a inespecífica, aumentó considerablemente a medida que los solicitantes disminuyeron las cantidades equimolares de SpCas9 y ARNsg transfectadas en células 293FT (Fig. 27c, d) de 7,1 x 10-10 a 1,8 x 10-11 nmol/célula (400 ng a 10 ng de plásmido Cas9-ARNsg). El ensayo qRT-PCR confirmó que el nivel de ARNm de hSpCas9 y ARNsg disminuyó proporcionalmente a la cantidad de ADN transfectado. Mientras que la especificidad aumentó gradualmente en casi 4 veces a medida que los solicitantes disminuyeron la cantidad de ADN transfectado de 7,1 x 10-10 a 9,0 x 10-11 nmol/célula (400 ng a 50 ng de plásmido), los solicitantes observaron un notable aumento adicional de 7 veces en la especificidad tras la disminución del ADN transfectado de 9,0 x 10-11 a 1,8 x 10-11 nmol/célula (50 ng a 10 ng de plásmido; Fig. 27c). Estos hallazgos sugieren que los solicitantes pueden minimizar el nivel de actividad inespecífica al valorar la cantidad de ADN de SpCas9 y ARNsg suministrados. Sin embargo, el aumento de la especificidad al reducir la cantidad de ADN transfectado también conduce a una reducción en la escisión en la diana. Estas mediciones permiten la integración cuantitativa de los criterios de especificidad y eficacia en la elección de dosis para optimizar la actividad de SpCas9 para diferentes aplicaciones. Los solicitantes exploran además las modificaciones en SpCas9 y el diseño de ARNsg que pueden mejorar la especificidad intrínseca sin sacrificar la eficacia de escisión. La figura 29 muestra los datos para la diana 2 y la diana 6 de EMX1. Para los sitios probados en las figuras 27 y 29 (en este caso, los sitios con 3 desapareamientos o menos), no se identificaron sitios inespecíficos (definidos como escisión del sitio inespecífica dentro de 100 veces de la escisión del sitio diana).

La capacidad de programar SpCas9 para dirigirse a sitios específicos en el genoma simplemente mediante el diseño de un ARNsg corto tiene un enorme potencial para varias aplicaciones. Los resultados de los solicitantes demuestran que la especificidad de la escisión del ADN mediada por SpCas9 depende de la secuencia y del locus y se rige por la cantidad, la posición y la identidad de las bases desapareadas. De manera importante, mientras que el PAM de 8-12 pb proximal de la secuencia guía generalmente define la especificidad, las secuencias PAM distales también contribuyen a la especificidad general de la escisión del ADN mediada por SpCas9. Aunque puede haber una escisión inespecífica para una secuencia de guía dada, pueden predecirse y probablemente minimizarse siguiendo las pautas generales de diseño.

Para maximizar la especificidad de SpCas9 para editar un gen en particular, se debe identificar posibles secuencias genómicas "inespecíficas" considerando las siguientes cuatro restricciones: En primer lugar y principalmente, no deben ir seguidos de un PAM con secuencias 5'-NGG o 5'-NAG. En segundo lugar, se debe minimizar su similitud de secuencia global con la secuencia diana, y se deben evitar las secuencias guía con loci genómicos inespecíficos que tienen menos de 3 desapareamientos. En tercer lugar, al menos 2 desapareamientos deben estar dentro de la región proximal de PAM del sitio inespecífico. En cuarto lugar, un número máximo de desapareamientos debe ser consecutivo o espaciado a menos de cuatro bases. Por último, la cantidad de SpCas9 y ARNsg se puede valorar para optimizar la relación de escisión específica a inespecífica.

Usando estos criterios, los solicitantes formularon un esquema de puntuación simple para integrar las contribuciones de la ubicación, densidad e identidad de desapareamientos para cuantificar su contribución al corte de SpCas9. Los solicitantes aplicaron las eficacias de escisión agregadas de los ARN guía de desapareamientos únicos para probar este esquema de puntuación por separado en dianas de todo el genoma. Los solicitantes encontraron que estos factores, considerados en su conjunto, representaron más del 50 % de la varianza en la clasificación de frecuencia de corte entre las dianas de todo el genoma estudiados (Fig. 30).

Aplicando las pautas delineadas anteriormente, los solicitantes diseñaron una herramienta computacional para facilitar la selección y validación de ARNsg, así como para predecir loci inespecíficos para análisis de especificidad; se puede

acceder a esta herramienta en el sitio web genome-engineering.org/tools. Estos resultados y herramientas amplían además el sistema SpCas9 como una alternativa poderosa y versátil a ZFN y TALEN para aplicaciones de edición de genomas. El trabajo adicional que examina la termodinámica y la estabilidad *in vivo* de los dúplex ARNsg-ADN probablemente proporcionará un poder predictivo adicional para la actividad inespecífica, mientras que la exploración de mutantes y ortólogos de SpCas9 puede producir nuevas variantes con una especificidad mejorada. Códigos de acceso Se puede acceder a todas las lecturas en bruto en NCB1 BioProject, número de referencia SRP023129.

Métodos y Materiales:

Cultivo celular y transfección: se mantuvo la línea celular 293FT de riñón embrionario humano (HEK) (Life Technologies) en medio de Eagle modificado por Dulbeeco (DMEM) complementado con suero bovino fetal al 10 % (HyClone), GlutaMAX 2 mM (Life Technologies), penicilina 100 U/ml y estreptomicina 100 µg/ml a 37 °C con incubación de CO2 al 5 %.

Se sembraron las células 293FT en placas de 6 pocillos, placas de 24 pocillos o placas de 96 pocillos (Corning) 24 horas antes de la transfección. Se transfectaron las células utilizando Lipofectamine 2000 (Life Technologies) en una confluencia del 80-90 % siguiendo el protocolo recomendado por el fabricante. Para cada pocillo de una placa de 6 pocillos, se utilizó un total de 1 µg de plásmido de Cas9+ARNsg. Para cada pocillo de una placa de 24 pocillos, se utilizó un total de 500 ng de plásmido de Cas9+ARNg a menos que se indique lo contrario. Para cada pocillo de una placa de 96 pocillos, se utilizaron 65 ng de plásmido Cas9 en una relación molar de 1:1 con respecto al producto de PCR de U6-ARNsg.

La línea de células madre embrionarias humanas HUES9 (núcleo del Instituto de células madre de Harvard) se mantuvo en condiciones libres de alimentador en GelTrex (Life Technologies) en medio mTesR (Stemcell Technologies) complementado con 100 µg/ml de Normocin (invivoGen). Las células HUES9 se transfectaron con el kit Nucleofector 4-D de células primarias Amaxa P3 (Lonza) siguiendo el protocolo del fabricante.

Ensayo de nucleasa SURVEYOR para la modificación del genoma Se transfectaron células 293FT con ADN plasmídico como se describe anteriormente. Se incubaron las células a 37 °C durante 72 horas después de la transfección antes de la extracción del ADN genómico. El ADN genómico se extrajo utilizando la solución de extracción de ADN QuickExtract (Epicentre) siguiendo el protocolo del fabricante. En resumen, se resuspendieron las células sedimentadas en solución QuickExtract y se incubaron a 65 °C durante 15 minutos y a 98 °C durante 10 minutos.

La región genómica que flanquea el sitio diana CRISPR para cada gen se amplificó por PCR (los cebadores se enumeran en la Tabla 2), y los productos se purificaron utilizando la columna QiaQuick Spin (Qiagen) siguiendo el protocolo del fabricante. Se mezclaron 400 ng de productos de PCR purificados con 2 µl de tampón de PCR Taq DNA Polimerasa 10X (Enzymatics) y agua ultrapura hasta un volumen final de 20 µl, y se sometieron a un proceso de recocido para permitir la formación del heterodúplex: 95 °C durante 10 min, 95 °C a 85 °C disminuyendo - 2 °C/s, 85 °C a 25 °C a - 0,25 °C/s, y 25 °C se mantienen durante 1 minuto. Después del recocido, se trataron los productos con SURVEYOR nucleasa y SURVEYOR potenciador S (Transgenomics) siguiendo el protocolo recomendado por el fabricante, y se analizaron en geles de poliacrilamida Novex TBE al 4-20 % (Life Technologies). Los geles se tiñeron con tinte de ADN S YBR Gold (Life Technologies) durante 30 minutos y se tomaron imágenes con un sistema de imágenes en gel Gel Doc (Bio-rad). La cuantificación se basó en intensidades de banda relativas.

Análisis de transferencia Northern de la expresión de ARNtracr en células humanas: se realizaron transferencias Northern como se describió previamente 1. En resumen, se calentaron los ARN a 95 °C durante 5 minutos antes de cargarlos en geles de poliacrilamida desnaturalizantes al 8 % (SequaGel, National Diagnostics). Posteriormente, se transfirió el ARN a una membrana Hybond N+ prehibridada (GE Healthcare) y se reticuló con reticulante UV Stratagene (Stratagene). Las sondas se marcaron con [gamma-32P] ATP (Perkin Elmer) con polinucleótido quinasa T4 (New England Biolabs). Después del lavado, se expuso la membrana al tamiz de fósforo durante una hora y se exploró con una placa de fósforo (Typhoon).

Secuenciación de bisulfito para evaluar el estado de metilación del ADN: se transfectaron las células HEK 293FT con Cas9 como se describió anteriormente. Se aisló el ADN genómico con el kit DNeasy Blood & Tissue (Qiagen) y el bisulfito se convirtió con el kit EZ DNA Methylation-Lightning (Zymo Research). La PCR con bisulfito se llevó a cabo utilizando KAPA2G Robust HotStart ADN polimerasa (KAPA Biosystems) con cebadores diseñados con el buscador de cebadores de bisulfito (Zymo Research, Tabla 6). Los amplicones de PCR resultantes se purificaron en gel, se digirieron con EcoRI y HindIII, y se ligaron en un esqueleto pUC19 antes de la transformación. Los clones individuales fueron luego secuenciados por Sanger para evaluar el estado de metilación del ADN.

Transcripción *in vitro* y ensayo de escisión: se transfectaron las células HEK 293FT con Cas9 como se describió anteriormente. Entonces se prepararon lisados celulares completos con un tampón de lisis (HEPES 20 mM, KCl 100 mM, MgCl2 5 mM, DTT 1 mM, glicerol al 5 %, Triton X-100 al 0,1%) complementado con un inhibidor de la proteasa Cocktail (Roche). El ARNsg dirigido por T7 se transcribió *in vitro* utilizando oligos personalizados (Sequences) y el kit de transcripción *in vitro* HiScribe T7 (NEB), siguiendo el protocolo recomendado por el fabricante. Para preparar sitios diana metilados, el plásmido pUC19 se metiló por M.SssI y luego se linealizó por NheI. El ensayo de escisión *in vitro*

se realizó de la siguiente manera: para una reacción de escisión de 20 µL, 10 µL de lisado celular se incuba con 2 µL de tampón de escisión (HEPES 100 mM, KCl 500 mM, MgCl2 25 mM, DTT 5 mM, glicerol al 25 %), el ARN transcrito *in vitro* y 300 ng de ADN plasmídico pUC19.

Secuenciación profunda para evaluar la especificidad dirigida: Se transfectaron células HEK 293FT colocadas en placas de 96 pocillos con ADN plasmídico Cas9 y casete de PCR de ARN guía único (ARNsg) 72 horas antes de la extracción del ADN genómico (Fig. 14). La región genómica que flanquea el sitio diana CRISPR para cada gen se amplificó mediante un método de PCR de fusión para unir los adaptadores Illumina P5, así como los códigos de barras específicos de la muestra únicos a los amplicones diana. Los productos de PCR se purificaron utilizando placas de filtro de 96 pocillos EconoSpin (Epoch Life Sciences) siguiendo el protocolo recomendado por el fabricante.

Las muestras de ADN con código de barras y purificadas se cuantificaron mediante el kit de análisis de dsDNA Quant-iT PicoGreen o el fluorómetro Qubit 2.0 (Life Technologies) y se agruparon en una proporción equimolar. Las bibliotecas de secuenciación se secuenciaron a continuación con el secuenciador personal Illumina MiSeq (Life Technologies).

Secuenciación de análisis de datos y detección de indel: Las lecturas de MiSeq se filtraron al exigir una calidad de Phred promedio (puntuación Q) de al menos 23, así como coincidencias de secuencia perfectas con los códigos de barras y los cebadores directos de amplicón. Las lecturas de los loci específicos e inespecíficos se analizaron realizando primero alineamientos de Smith-Waterman contra secuencias de amplicón que incluían 50 nucleótidos en cadena arriba y cadena abajo del sitio diana (un total de 120 pb). Mientras tanto, los alineamientos se analizaron para detectar indeles desde 5 nucleótidos cadena arriba hasta 5 nucleótidos cadena abajo del sitio diana (un total de 30 pb). Las regiones diana analizadas se descartaron si parte de su alineamiento se encontraba fuera del MiSeq leído, o si los pares de bases apareados comprendían menos del 85 % de su longitud total.

Los controles negativos para cada muestra proporcionaron un indicador para la inclusión o exclusión de los indeles como supuestos eventos de corte. Para cada muestra, solo se contó un indel si su puntuación de calidad superó a μ-σ, donde μ fue la puntuación de calidad media del control negativo correspondiente a esa muestra y σ fue la desviación estándar de la misma. Esto produjo tasas de indel de región diana completa tanto para los controles negativos como para sus muestras correspondientes. Utilizando la tasa de error del control negativo por región diana por lectura, *q,* el recuento de indeles observado de la muestra *n*, y su recuento de lecturas *R*, una estimación de probabilidad máxima para la fracción de lecturas que tienen regiones diana con índices verdaderos, *p,* se derivó mediante la aplicación de un modelo de error binomial, del siguiente modo.

Dejando que el número (desconocido) de lecturas en una muestra que tiene regiones diana incorrectamente contadas como que tiene al menos 1 indel sea *E*, los solicitantes pueden escribir (sin hacer ninguna suposición sobre el número de indeles verdaderos)

$$\mathrm{Prob}(E|p) = \binom{R(1-p)}{E} q^E (1-q)^{R(1-p)-E}$$

ya que $R(1-p)$ es el número de lecturas que tienen regiones diana sin indeles verdaderos. Paralelamente, debido a que el número de lecturas que se observan con indeles es n, $n = E + Rp$, en otras palabras, el número de lecturas que tiene regiones diana con errores pero sin indeles verdaderos más el número de lecturas cuyas regiones diana tienen correctamente indeles. Los solicitantes pueden volver a escribir lo anterior

$$\mathrm{Prob}(E|p) = \mathrm{Prob}(n = E + Rp|p) = \binom{R(1-p)}{n-Rp} q^{n-Rp} (1-q)^{E-n}$$

Tomando todos los valores de la frecuencia de las regiones diana con los indeles verdaderos p para que sean igualmente probables a priori, Prob$(n|p) \propto$ Prob$(p|n)$. Por lo tanto, la estimación de máxima verosimilitud (MLE) para la frecuencia de las regiones diana con los verdaderos indeles se estableció como el valor de p que maximizó el Prob $(n|p)$. Esto fue evaluado numéricamente.

Para colocar los límites de error en las frecuencias de lectura de verdadero indel en las propias bibliotecas de secuenciación, se calcularon los intervalos de puntuación de Wilson (2) para cada muestra, dada la estimación de MLE para regiones diana de verdadero indel, *Rp,* y el número de lecturas *R.* Explícitamente, el límite inferior *l* y el límite superior *u* se calcularon como

$$l = \left(Rp + \frac{z^2}{2} - z\sqrt{Rp(1-p) + z^2/4} \right) / (R + z^2)$$

$$u = \left(Rp + \frac{z^2}{2} + z\sqrt{Rp(1-p) + z^2/4} \right) / (R + z^2)$$

donde z, la puntuación estándar para la confianza requerida en la distribución normal de la varianza 1, se estableció en 1,96, lo que significa una confianza del 95 %.

Análisis qRT-PCR de la expresión relativa de Cas9 y ARNsg: Se transfectaron células 293FT colocadas en placas de 24 pocillos como se describe anteriormente. A las 72 horas de la transfección, se recogió el ARN total con el kit miRNeasy Micro (Qiagen). La síntesis de cadena inversa para ARNsg se realizó con el kit de ADNc Flex qScript (VWR) y los cebadores de síntesis de primera cadena personalizados (Tabla 6). El análisis de qPCR se realizó con Fast SYBR Green Master Mix (Life Technologies) y cebadores personalizados (Tabla 2), utilizando GAPDH como control endógeno. La cuantificación relativa se calculó mediante el método ΔΔCT.

Tabla 1 | Secuencias de sitios diana. Los sitios diana probados para el sistema CRISPR tipo II de *S. pyogenes* con el PAM requerido. Las células se transfectaron con Cas9 y con ARNcr-ARNtracr o ARNsg quimérico para cada diana.

ID sitio diana	diana genómica	Secuencia de sitio diana (5' a 3')	PAM	cadena
1	*EMX1*	GTCACCTCCAATGACTAGGG	TGG	+
2	*EMX1*	GACATCGATGTCCTCCCCAT	TGG	-
3	*EMX1*	GAGTCCGAGCAGAAGAAGAA	GGG	+
6	*EMX1*	GCGCCACCGGTTGATGTGAT	GGG	-
10	*EMX1*	GGGGCACAGATGAGAAACTC	AGG	-
11	*EMX1*	GTACAAACGGCAGAAGCTGG	AGG	+
12	*EMX1*	GGCAGAAGCTGGAGGAGGAA	GGG	+
13	*EMX1*	GGAGCCCTTCTTCTTCTGCT	CGG	-
14	*EMX1*	GGGCAACCACAAACCCACGA	GGG	+
15	*EMX1*	GCTCCCATCACATCAACCGG	TGG	+
16	*EMX1*	GTGGCGCATTGCCACGAAGC	AGG	+
17	*EMX1*	GGCAGAGTGCTGCTTGCTGC	TGG	+
18	*EMX1*	GCCCCTGCGTGGGCCCAAGC	TGG	+
19	*EMX1*	GAGTGGCCAGAGTCCAGCTT	GGG	-
20	*EMX1*	GGCCTCCCCAAAGCCTGGCC	AGG	-
4	*PVALB*	GGGGCCGAGATTGGGTGTTC	AGG	+
5	*PVALB*	GTGGCGAGAGGGGCCGAGAT	TGG	+
1	*SERPINB5*	GAGTGCCGCCGAGGCGGGGC	GGG	+
2	*SERPINB5*	GGAGTGCCGCCGAGGCGGGG	CGG	+
3	*SERPINB5*	GGAGAGGAGTGCCGCCGAGG	CGG	+

Tabla 2 | Secuencias de cebadores

Ensayo SURVEYOR		
nombre del cebador	**diana genómica**	**secuencia del cebador (5' a 3')**
Sp-EMX1-F1	*EMX1*	AAAACCACCCTTCTCTCTGGC

Sp-EMX1-R1	*EMX1*	GGAGATTGGAGACACGGAGAG
Sp-EMX1-F2	*EMX1*	CCATCCCCTTCTGTGAATGT
Sp-EMX1-R2	*EMX1*	GGAGATTGGAGACACGGAGA
Sp-PVALB-F	*PVALB*	CTGGAAAGCCAATGCCTGAC
Sp-PVALB-R	*PVALB*	GGCAGCAAACTCCTTGTCCT
qRT-PCR de la expresión de Cas9 y ARNsg		
nombre del cebador	**secuencia del cebador (5' a 3')**	
síntesis de ARNsg de cadena inversa	AAGCACCGACTCGGTGCCAC	
EMX1.1 ARNsg qPCR F	TCACCTCCAATGACTAGGGG	
EMX1.1 ARNsg qPCR R	CAAGTTGATAACGGACTAGCCT	
EMX1.3 ARNsg qPCR F	AGTCCGAGCAGAAGAAGAAGTTT	
EMX1.3 ARNsg qPCR R	TTTCAAGTTGATAACGGACTAGCCT	
Cas9 qPCR F	AAACAGCAGATTCGCCTGGA	
Cas9 qPCR R	TCATCCGCTCGATGAAGCTC	
GAPDH qPCR F	TCCAAAATCAAGTGGGGCGA	
GAPDH qPCR R	TGATGACCCTTTTGGCTCCC	
PCR y secuenciación de bisulfito		
nombre del cebador	**secuencia del cebador (5' a 3')**	
PCR D de bisulfito (*locus SERPINB5*)	GAGGAATTCTTTTTTTGTTYGAATATGTTGGAGGT TTTTTGGAAG	
PCR I con bisulfito (locus *SERPINB5*)	GAGAAGCTTAAATAAAAAACRACAATACTCAACC CAACAACC	
Secuenciación de pUC19	CAGGAAACAGCTATGAC	

Tabla 3 | Secuencias de cebadores para probar la arquitectura del ARNsg. Los cebadores se hibridan con la cadena inversa del promotor U6 a menos que se indique lo contrario. El sitio de cebado U6 está en negrita, la secuencia guía está indicada por el tramo de "N", la secuencia de repetición directa está en cursiva y la secuencia ARNtracr está subrayada. La estructura secundaria de cada arquitectura del ARNsg se muestra en la figura 71.

nombre del cebador	**secuencia del cebador (5' a 3')**
Directo de U6	GCCTCTA**GAGGTACCTGAGGGCCTATTTCCCATGATTCC**
I: ARNsg (DR +12, ARNtracr -+-85)	ACCTCTAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATTTTAACTTGCTATTTC*TAGCTCTA AAAC*CNNNNNNNNNNNNNNNNNNN**GGTGTTTCGTCCTTTCC ACAAG**
II: ARNsgCDR 12. ARNtracr -+85) mut2	ACCTCTAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATATTAACTTGCTATTTC*TAGCTCTA ATAC*CNNNNNNNNNNNNNNNNNNN**GGTGTTTCGTCCTTTCC ACAAG**
III: ARNsg (DR +22, ARNtracr -+-85)	ACCTCTAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGTTTTGTT TC*CAAAACAGCATAGCTCTAAAAC*NNNNNNNNNNNNNNNNNN **NGGTGTTTCGTCCTTTCCACAAG**

IV: ARNsg(DR +22, ARNtracr +85) mut4	ACCTCTAGAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGT TGATAACGGACTAGCCTTATATTAACTTGCTATGCTGTATTGT TTC*CAATACAGCATAGCTCTAATAC*NNNNNNNNNNNNNNNNN NNGGTGTTTCGTCCTTTCCACAAG

Tabla 4 | Sitios diana con PAM alternativos para probar la especificidad de PAM de Cas9. Todos los sitios diana para las pruebas de especificidad de PAM se encuentran dentro del locus EMX1 humano.

Secuencia del sitio diana (5' a 3')	PAM
AGGCCCCAGTGGCTGCTCT	NAA
ACATCAACCGGTGGCGCAT	NAT
AAGGTGTGGTTCCAGAACC	NAC
CCATCACATCAACCGGTGG	NAG
AAACGGCAGAAGCTGGAGG	NTA
GGCAGAAGCTGGAGGAGGA	NTT
GGTGTGGTTCCAGAACCGG	NTC
AACCGGAGGACAAAGTACA	NTG
TTCCAGAACCGGAGGACAA	NCA
GTGTGGTTCCAGAACCGGA	NCT
TCCAGAACCGGAGGACAAA	NCC
CAGAAGCTGGAGGAGGAAG	NCG
CATCAACCGGTGGCGCATT	NGA
GCAGAAGCTGGAGGAGGAA	NGT
CCTCCCTCCCTGGCCCAGG	NGC
TCATCTGTGCCCCTCCCTC	NAA
GGGAGGACATCGATGTCAC	NAT
CAAACGGCAGAAGCTGGAG	NAC
GGGTGGGCAACCACAAACC	NAG
GGTGGGCAACCACAAACCC	NTA
GGCTCCCATCACATCAACC	NTT
GAAGGGCCTGAGTCCGAGC	NTC
CAACCGGTGGCGCATTGCC	NTG
AGGAGGAAGGGCCTGAGTC	NCA
AGCTGGAGGAGGAAGGGCC	NCT
GCATTGCCACGAAGCAGGC	NCC
ATTGCCACGAAGCAGGCCA	NCG
AGAACCGGAGGACAAAGTA	NGA
TCAACCGGTGGCGCATTGC	NGT
GAAGCTGGAGGAGGAAGGG	NGC

SECUENCIAS

Todas las secuencias están en la dirección 5' a 3'. Para la transcripción de U6, la cadena de Ts subrayada sirve como terminador de la transcripción.

> U6-ARNtracr corto (*Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccGGAACCATTCAAAACAGCATAGCA
AGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGT
GCTTTTTTT

(la secuencia del ARNtracr está en negrita)
> U6-DR-secuencia guía-DR. *(Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccg*gttttagagctatgctgttttgaatggtcccaaaac*NN
NNNNNNNNNNNNNNNNNNNNNNNNNNNN*gttttagagctatgctgttttgaatggtcccaaaac*TTTTTT
T

la secuencia de repetición directa está en cursiva y la secuencia guía está indicada por el tramo de "N")
> ARNsg que contiene +48 ARNtracr (*Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNN*gttttaga*
*gcta*gaaa**tagcaagttaaaataaggctagtccg**TTTTTTT

(la secuencia guía se resalta en azul y el fragmento ARNtracr está en negrita)
> ARNsg que contiene +54 ARNtracr (*Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNN*gttttaga*
*gcta*gaaa**tagcaagttaaaataaggctagtccgttatca**TTTTTTTT

(la secuencia guía está indicada por el tramo de "N" s y el fragmento ARNtracr está en negrita)
> ARNsg que contiene +67 ARNtracr (*Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNN*gttttaga*
*gcta*gaaa**tagcaagttaaaataaggctagtccgttatcaacttgaaaaagtg**TTTTTTT

(la secuencia guía está indicada por el tramo de "N" y el fragmento ARNtracr está en negrita))
> ARNsg que contiene +85 ARNtracr (*Streptococcus pyogenes* SF370)

gagggcctatttcccatgattccttcatatttgcatatacgatacaaggctgttagagagataattggaattaatttgactgtaaacacaaagatat
tagtacaaaatacgtgacgtagaaagtaataatttcttgggtagtttgcagttttaaaattatgttttaaaatggactatcatatgcttaccgtaactt
gaaagtatttcgatttcttggctttatatatcttgtggaaaggacgaaacaccNNNNNNNNNNNNNNNNNNNN*gttttaga*
*gcta*gaaa**tagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgc**TTTTTTT

(la secuencia guía está indicada por el tramo de "N" s y el fragmento ARNtracr está en negrita)
> CBh-NLS-SCas9-NLS

CGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCC
ATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTG
ACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTA
TCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCA
TTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTA
GTCATCGCTATTACCATGGTCGAGGTGAGCCCCACGTTCTGCTTCACTCTCCCCATCT
CCCCCCCCCTCCCCACCCCCAATTTTGTATTTATTTATTTTTTAATTATTTTGTGCAGCG
ATGGGGGCGGGGGGGGGGGGGGGGCGCGCGCCAGGCGGGGCGGGGCGGGGCGAG
GGGCGGGGCGGGGCGAGGCGGAGAGGTGCGGCGGCAGCCAATCAGAGCGGCGCGC
TCCGAAAGTTTCCTTTTATGGCGAGGCGGCGGCGGCGGCGGCCCTATAAAAAGCGA
AGCGCGCGGCGGGCGGGAGTCGCTGCGACGCTGCCTTCGCCCCGTGCCCCGCTCCG
CCGCCGCCTCGCGCCGCCCGCCCCGGCTCTGACTGACCGCGTTACTCCCACAGGTGA
GCGGGCGGGACGGCCCTTCTCCTCCGGGCTGTAATTAGCTGAGCAAGAGGTAAGGG
TTTAAGGGATGGTTGGTTGGTGGGGTATTAATGTTTAATTACCTGGAGCACCTGCCT
GAAATCACTTTTTTTCAGGTTGGaccggtgccacc**ATGGACTATAAGGACCACGACGGAG**

ACTACAAGGATCATGATATTGATTACAAAGACGATGACGATAAGATGGCCCCA
AAGAAGAAGCGGAAGGTCGGTATCCACGGAGTCCCAGCAGCCGACAAGAAGTA
CAGCATCGGCCTGGACATCGGCACCAACTCTGTGGGCTGGGCCGTGATCACCG
ACGAGTACAAGGTGCCCAGCAAGAAATTCAAGGTGCTGGGCAACACCGACCGG
CACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAAC
AGCCGAGGCCACCCGGCTGAAGAGAACCGCCAGAAGAAGATACACCAGACGG
AAGAACCGGATCTGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGT
GGACGACAGCTTCTTCCACAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATA
AGAAGCACGAGCGGCACCCCATCTTCGGCAACATCGTGGACGAGGTGGCCTAC
CACGAGAAGTACCCCACCATCTACCACCTGAGAAAGAAACTGGTGGACAGCAC
CGACAAGGCCGACCTGCGGCTGATCTATCTGGCCCTGGCCCACATGATCAAGT
TCCGGGGCCACTTCCTGATCGAGGGCGACCTGAACCCCGACAACAGCGACGTG
GACAAGCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAGGAAAA
CCCCATCAACGCCAGCGGCGTGGACGCCAAGGCCATCCTGTCTGCCAGACTGA
GCAAGAGCAGACGGCTGGAAAATCTGATCGCCCAGCTGCCCGGCGAGAAGAA
GAATGGCCTGTTCGGCAACCTGATTGCCCTGAGCCTGGGCCTGACCCCCAACT
TCAAGAGCAACTTCGACCTGGCCGAGGATGCCAAACTGCAGCTGAGCAAGGAC
ACCTACGACGACGACCTGGACAACCTGCTGGCCCAGATCGGCGACCAGTACGC
CGACCTGTTTCTGGCCGCCAAGAACCTGTCCGACGCCATCCTGCTGAGCGACA
TCCTGAGAGTGAACACCGAGATCACCAAGGCCCCCCTGAGCGCCTCTATGATC
AAGAGATACGACGAGCACCACCAGGACCTGACCCTGCTGAAAGCTCTCGTGCG
GCAGCAGCTGCCTGAGAAGTACAAAGAGATTTTCTTCGACCAGAGCAAGAACG
GCTACGCCGGCTACATTGACGGCGGAGCCAGCCAGGAAGAGTTCTACAAGTTC
ATCAAGCCCATCCTGGAAAAGATGGACGGCACCGAGGAACTGCTCGTGAAGCT
GAACAGAGAGGACCTGCTGCGGAAGCAGCGGACCTTCGACAACGGCAGCATCC
CCCACCAGATCCACCTGGGAGAGCTGCACGCCATTCTGCGGCGGCAGGAAGAT
TTTTACCCATTCCTGAAGGACAACCGGGAAAAGATCGAGAAGATCCTGACCTTC
CGCATCCCCTACTACGTGGGCCCTCTGGCCAGGGGAAACAGCAGATTCGCCTG
GATGACCAGAAAGAGCGAGGAAACCATCACCCCCTGGAACTTCGAGGAAGTGG
TGGACAAGGGCGCTTCCGCCCAGAGCTTCATCGAGCGGATGACCAACTTCGAT
AAGAACCTGCCCAACGAGAAGGTGCTGCCCAAGCACAGCCTGCTGTACGAGTA
CTTCACCGTGTATAACGAGCTGACCAAAGTGAAATACGTGACCGAGGGAATGA
GAAAGCCCGCCTTCCTGAGCGGCGAGCAGAAAAAGGCCATCGTGGACCTGCTG
TTCAAGACCAACCGGAAAGTGACCGTGAAGCAGCTGAAAGAGGACTACTTCAA
GAAAATCGAGTGCTTCGACTCCGTGGAAATCTCCGGCGTGGAAGATCGGTTCA
ACGCCTCCCTGGGCACATACCACGATCTGCTGAAAATTATCAAGGACAAGGAC
TTCCTGGACAATGAGGAAAACGAGGACATTCTGGAAGATATCGTGCTGACCCT
GACACTGTTTGAGGACAGAGAGATGATCGAGGAACGGCTGAAAACCTATGCCC
ACCTGTTCGACGACAAAGTGATGAAGCAGCTGAAGCGGCGGAGATACACCGGC
TGGGGCAGGCTGAGCCGGAAGCTGATCAACGGCATCCGGGACAAGCAGTCCG
GCAAGACAATCCTGGATTTCCTGAAGTCCGACGGCTTCGCCAACAGAAACTTC
ATGCAGCTGATCCACGACGACAGCCTGACCTTTAAAGAGGACATCCAGAAAGC
CCAGGTGTCCGGCCAGGGCGATAGCCTGCACGAGCACATTGCCAATCTGGCCG
GCAGCCCCGCCATTAAGAAGGGCATCCTGCAGACAGTGAAGGTGGTGGACGAG
CTCGTGAAAGTGATGGGCCGGCACAAGCCCGAGAACATCGTGATCGAAATGGC
CAGAGAGAACCAGACCACCCAGAAGGGACAGAAGAACAGCCGCGAGAGAATG

AAGCGGATCGAAGAGGGCATCAAAGAGCTGGGCAGCCAGATCCTGAAAGAACA
CCCCGTGGAAAACACCCAGCTGCAGAACGAGAAGCTGTACCTGTACTACCTGC
AGAATGGGCGGGATATGTACGTGGACCAGGAACTGGACATCAACCGGCTGTCC
GACTACGATGTGGACCATATCGTGCCTCAGAGCTTTCTGAAGGACGACTCCAT
CGACAACAAGGTGCTGACCAGAAGCGACAAGAACCGGGGCAAGAGCGACAAC
GTGCCCTCCGAAGAGGTCGTGAAGAAGATGAAGAACTACTGGCGGCAGCTGCT
GAACGCCAAGCTGATTACCCAGAGAAAGTTCGACAATCTGACCAAGGCCGAGA
GAGGCGGCCTGAGCGAACTGGATAAGGCCGGCTTCATCAAGAGACAGCTGGTG
GAAACCCGGCAGATCACAAAGCACGTGGCACAGATCCTGGACTCCCGGATGAA
CACTAAGTACGACGAGAATGACAAGCTGATCCGGGAAGTGAAAGTGATCACCC
TGAAGTCCAAGCTGGTGTCCGATTTCCGGAAGGATTTCCAGTTTTACAAAGTGC
GCGAGATCAACAACTACCACCACGCCCACGACGCCTACCTGAACGCCGTCGTG
GGAACCGCCCTGATCAAAAAGTACCCTAAGCTGGAAAGCGAGTTCGTGTACGG
CGACTACAAGGTGTACGACGTGCGGAAGATGATCGCCAAGAGCGAGCAGGAAA
TCGGCAAGGCTACCGCCAAGTACTTCTTCTACAGCAACATCATGAACTTTTTCA
AGACCGAGATTACCCTGGCCAACGGCGAGATCCGGAAGCGGCCTCTGATCGAG
ACAAACGGCGAAACCGGGGAGATCGTGTGGGATAAGGGCCGGGATTTTGCCAC
CGTGCGGAAAGTGCTGAGCATGCCCCAAGTGAATATCGTGAAAAAGACCGAGG
TGCAGACAGGCGGCTTCAGCAAAGAGTCTATCCTGCCCAAGAGGAACAGCGAT
AAGCTGATCGCCAGAAAGAAGGACTGGGACCCTAAGAAGTACGGCGGCTTCGA
CAGCCCCACCGTGGCCTATTCTGTGCTGGTGGTGGCCAAAGTGGAAAAGGGCA
AGTCCAAGAAACTGAAGAGTGTGAAAGAGCTGCTGGGGATCACCATCATGGAA
AGAAGCAGCTTCGAGAAGAATCCCATCGACTTTCTGGAAGCCAAGGGCTACAA
AGAAGTGAAAAAGGACCTGATCATCAAGCTGCCTAAGTACTCCCTGTTCGAGC
TGGAAAACGGCCGGAAGAGAATGCTGGCCTCTGCCGGCGAACTGCAGAAGGG
AAACGAACTGGCCCTGCCCTCCAAATATGTGAACTTCCTGTACCTGGCCAGCCA
CTATGAGAAGCTGAAGGGCTCCCCCGAGGATAATGAGCAGAAACAGCTGTTTG
TGGAACAGCACAAGCACTACCTGGACGAGATCATCGAGCAGATCAGCGAGTTC
TCCAAGAGAGTGATCCTGGCCGACGCTAATCTGGACAAAGTGCTGTCCGCCTA
CAACAAGCACCGGGATAAGCCCATCAGAGAGCAGGCCGAGAATATCATCCACC
TGTTTACCCTGACCAATCTGGGAGCCCCTGCCGCCTTCAAGTACTTTGACACCA
CCATCGACCGGAAGAGGTACACCAGCACCAAAGAGGTGCTGGACGCCACCCTG
ATCCACCAGAGCATCACCGGCCTGTACGAGACACGGATCGACCTGTCTCAGCT
GGGAGGCGACTTTCTTTTTCTTAGCTTGACCAGCTTTCTTAGTAGCAGCAGGAC
GCTTTAA

(NLS-hSpCas9-NLS está en negrita)

> Amplicón de secuenciación para las guías de EMX1 1.1, 1.14, 1.17

CCAATGGGGAGGACATCGATGTCACCTCCAATGACTAGGGTGGGCAACCACAAACC
CACGAGGGCAGAGTGCTGCTTGCTGCTGGCCAGGCCCCTGCGTGGGCCCAAGCTGG
ACTCTGGCCAC

> Amplicón de secuenciación para las guías de EMX1 1.2, 1.16

CGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGC
AGGCCAATGGGGAGGACATCGATGTCACCTCCAATGACTAGGGTGGGCAACCACAA
ACCCACGAG

> Amplicón de secuenciación para las guías de EMX1 1.3, 1.13, 1.15

GGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCA
GAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGCAGGCCA
ATGGGGAGGACATCGAT

> Amplicón de secuenciación para las guías de EMX1 1.6

AGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCA
CATCAACCGGTGGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCGATGTCA
CCTCCAATGACTAGGGTGG

> Amplicón de secuenciación para las guías de EMX1 1.10

CCTCAGTCTTCCCATCAGGCTCTCAGCTCAGCCTGAGTGTTGAGGCCCCAGTGGCTG
CTCTGGGGGCCTCCTGAGTTTCTCATCTGTGCCCCTCCCTCCCTGGCCCAGGTGAAG
GTGTGGTTCCA

> Amplicón de secuenciación para las guías de EMX1 1.11, 1.12

TCATCTGTGCCCCTCCCTCCCTGGCCCAGGTGAAGGTGTGGTTCCAGAACCGGAGGA
CAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAA
GAAGGGCTCCCATCACA

> Amplicón de secuenciación para las guías de EMX1 1.18, 1.19

CTCCAATGACTAGGGTGGGCAACCACAAACCCACGAGGGCAGAGTGCTGCTTGCTG
CTGGCCAGGCCCCTGCGTGGGCCCAAGCTGGACTCTGGCCACTCCCTGGCCAGGCTT
TGGGGAGGCCTGGAGT

> Amplicón de secuenciación para las guías de EMX1 1.20

CTGCTTGCTGCTGGCCAGGCCCCTGCGTGGGCCCAAGCTGGACTCTGGCCACTCCCT
GGCCAGGCTTTGGGGAGGCCTGGAGTCATGGCCCCACAGGGCTTGAAGCCCGGGGC
CGCCATTGACAGAG

> cebador D del promotor T7 para el recocido con la cadena diana

GAAATTAATACGACTCACTATAGGG

> oligo que contiene el sitio diana pUC19 1 para la metilación (T7 inverso)

AAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTT
AACTTGCTATTTCTAGCTCTAAAACAACGACGAGCGTGACACCACCCTATAGTGAGT
CGTATTAATTTC

> oligo que contiene el sitio diana pUC19 2 para la metilación (T7 inverso)

AAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTT
AACTTGCTATTTCTAGCTCTAAAACGCAACAATTAATAGACTGGACCTATAGTGAGT
CGTATTAATTTC

Ejemplo 7: Investigaciones de desapareamiento con el par de bases

Los solicitantes han probado varias secuencias de guía de desapareamientos que se han probado para probar el efecto inespecífico de la proteína cas9 (Hsu et al., DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013 Sep;31(9):827-32). Sin embargo, el número de secuencias de guía de desapareamientos utilizadas puede haber sido mucho menor que el conjunto completo de todas las posibles secuencias de guía de desapareamientos. Para predecir la eficacia de corte inespecífica de los pares guía-diana no probados, se construye un modelo termodinámico para ajustarse a los datos experimentales existentes y se usa para predecir el nuevo par diana guía.

Los solicitantes probaron los datos de frecuencia de corte para el par diana guía analizada recopilada de Hsu P.D et

al., 2013. Los solicitantes recopilaron los parámetros termodinámicos del vecino más cercano para el emparejamiento de los ácidos nucleicos de la distribución del paquete ViennaRNA. Los solicitantes obtuvieron los primeros parámetros dinámicos del par bi-base por posición para cada par diana guía. Estos parámetros fueron referidos como características. Los datos, que comprenden la frecuencia de corte y las características para cada par diana guía, se dividieron aleatoriamente en 5 grupos para realizar una validación cruzada de 5 veces. En cada sesión de capacitación, se utilizaron 4 grupos de datos como datos de capacitación. Se usó un núcleo de base radial modificado que consistía de un núcleo de base radial con un término aditivo que explicaba el PAM diferente en la regresión no lineal. En cada sesión de prueba, el 1 grupo restante se utilizó para probar el modelo. Se realizó una búsqueda de cuadrícula gruesa para los parámetros de ajuste. Los mejores parámetros de ajuste se utilizaron para generar nuevos parámetros en la búsqueda de cuadrícula fina.

El algoritmo de los solicitantes fluye de la siguiente manera:

1. Elegir los datos con duplicado.
2. Calcular el intervalo de confianza para la frecuencia de corte basado en la distribución binomial. Seleccionar datos con valor p pasando un umbral.
3. Calcular los parámetros dinámicos térmicos para cada par guía-diana para obtener su característica.
4. Transformar los datos de frecuencia de corte no lineal para equilibrar las contribuciones de frecuencia de corte alta y baja en el ajuste del modelo.
5. Dividir los datos en 5 grupos.
6. Realizar una búsqueda de cuadrícula gruesa para el parámetro de ajuste utilizando la validación cruzada de 5 veces.
7. Se seleccionan los mejores parámetros y se realiza la búsqueda de cuadrícula fina.

Resultado: Los datos de capacitación y los datos de prueba se muestran como puntos azules y puntos rojos, respectivamente, en el diagrama de dispersión. El coeficiente de correlación de Spearman es 0,88 y el coeficiente de correlación de Pearson es 0,91 (Fig. 36). Esto demuestra buen modelo de predicción de los datos. Para verificar el ajuste del modelo, los datos se aleatorizaron nuevamente, se captaron y se probaron con los mismos parámetros. El resultado es 0,82 para ambos coeficientes de correlación (Fig. 37).

REFERENCIAS

1. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339,819-823(2013)
2. Mali, P. et al. RNA-Guided Human Genome Engineering via Cas9. Science 339, 823-826 (2013).
3. Jinek, M. et al. RNA-programmed genome editing in human cells, eLife 2, e00471 (2013).
4. Cho, S.W., Kim, S., Kim, J.M. & Kim, J.S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol 31, 230-232 (2013).
5. Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature 471, 602-607 (2011).
6. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816-821 (2012).
7. Wang, H. et al. One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. Cell 153, 910-918 (2013).
8. Guschin, D.Y. et al. A rapid and general assay for monitoring endogenous gene modification. Methods Mol Biol 649, 247-256 (2010).
9. Bogenhagen, D.F. & Brown, D.D, Nucleotide sequences in Xenopus 5S DNA required for transcription termination. Cell 24, 261-270 (1981).
10. Hwang, W.Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol 31, 227-229 (2013).
11. Bultmann, S. et al. Targeted transcriptional activation of silent oct.4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. Nucleic Acids Res 40, 5368-5377 (2012),
12. Valton, J. et al. Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. J Biol Chem 287, 38427-38432 (2012).
13. Christian, M, et al. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics 186, 757-761 (2010).
14. Miller, J.C. et al. A TALE nuclease architecture for efficient genome editing. Nat Biotechnol 29, 143-148 (2011).
15. Mussolino, C. et al. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic acids research 39, 9283-9293 (2011).
16. Hsu, P.D. & Zhang, F. Dissecting neural function using targeted genome engineering technologies. ACS chemical neuroscience 3, 603-610 (2012).
17. Sanjana, N.E. et al. A transcription activator-like effector toolbox for genome engineering. Nature protocols 7, 171-192 (2012).
18. Porteus, M.H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. Science 300, 763 (2003).
19. Miller, J.C. et al. An improved zinc-finger nuclease architecture for highly specific genome editing, Nat Biotechnol 25, 778-785 (2007).

20. Sander, J.D. et al. Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA). Nat Methods 8, 67-69 (2011).
21. Wood, A.J. et al. Targeted genome editing across species using ZFNs and TALENs. Science 333, 307(2011).
22. Bobis-Wozowicz, S., Osiak, A., Rahman, S.H. & Cathomen, T. Targeted genome editing in pluripotent stem cells using zinc-finger nucleases. Methods 53, 339-346 (2011).
23. Jiang, W., Bikard, D., Cox, D., Zhang, F, & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol 31, 233-239 (2013).
24. Qi, L.S. et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell 152, 1173-1183 (2013).
25. Michaelis, L.M., Maud "Die kinetik der invertinwirkung.". Biochem. z (1913).
26. Mahfouz, M.M. et al. De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. Proc Natl Acad Sci U S A 108, 2623-2628 (2011).
27. Wilson, E.B, Probable inference, the law of succession, and statistical inference. .1 Am Stat Assoc 22, 209-212 (1927).
28. Ding, Q. et al. A. TALEN genome-editing system for generating human stem cell-based disease models. Cell Stem Cell 12, 238-251 (2013).
29. Soldner, F. et al. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. Cell 146, 318-331 (2011).
30. Carlson, D.F. et al. Efficient TALEN-mediated gene knockout in livestock. Proc Natl Acad Sci U S A 109, 17382-17387 (2012).
31. Geurts, A.M. et al. Knockout Rats via Embryo Microinjection of Zinc-Finger Nucleases. Science 325, 433-433 (2009).
32. Takasu, Y. et al. Targeted mutagenesis in the silkworm Bombyx mori using zinc finger nuclease mRNA injection. Insect Biochem Molec 40, 759-765 (2010).
33. Watanabe, T. et al. Non-transgenic genome modifications in a hemimetabolous insect using zinc-finger and TAL effector nucleases. Nat Commun 3 (2012).
34. Reyon, D. et al. FLASH assembly of TALENs for high-through put genome editing, Nat Biotechnol 30, 460-465 (2012).
35. Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326, 1509-1512 (2009).
36. Moscou, M.J, & Bogdanove, A.J. A simple cipher governs DNA recognition by TAL effectors. Science 326, 1501 (2009).
37. Deveau, H., Garneau. J.E. & Moineau, S. CRISPR/Cas system and its role in phage-bacteria interactions. Annu Rev Microbiol 64, 475-493 (2010).
38. Horvath. P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. Science 327, 167-170 (2010).
39. Makarova, K.S. et al. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9, 467-477 (2011).
40. Bhaya, D., Davison, M. & Barrangou, R. CRISPR-Cas systems in bacteria and archaea; versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 45, 273-297 (2011).
41. Gameau, J.E, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468, 67-71 (2010).
42. Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V, Cas9-ARNcr ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A 109, E2579-2586 (2012).
43. Urnov, F. D., Rebar, E.J., Holmes, MC, Zhang, H.S. & Gregory, P.D. Genome editing with engineered zinc finger nucleases. Nat Rev Genet 11, 636-646 (2010).
44. Perez, E.E. et al. Establishment of HIV-1 resistance in CD4(+) T cells by genome editing using zinc-finger nucleases. Nat Biotechnol 26, 808-816 (2008).
45. Chen, F.Q. et al. High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. Nat Methods 8, 753-U796 (2011).
46. Bedell, V.M. et al. In vivo genome editing using a high-efficiency TALEN system. Nature 491, 114-U133 (2012).
47. Saleh-Gohari, N. & Helleday, T. Conservative homologous recombination preferentially repairs DNA doublestrand breaks in the S phase of the cell cycle in human cells. Nucleic Acids Res 32, 3683-3688 (2004).
48. Sapranauskas, R. et al. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res 39, 9275-9282 (2011).
49. Shen, B. et al. Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. Cell Res 23, 720-723 (2013).
50. Tuschl, T. Expanding small RNA interference. Nat Biotechnol 20, 446-448 (2002).
51. Smithies, O., Gregg, R.G., Boggs, S.S., Koralewski, M.A. & Kucherlapati, R.S. Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. Nature 317, 230-234 (1985).
52. Thomas, K.R., Folger, K.R. & Capecchi, M.R. High frequency targeting of genes to specific sites in the mammalian genome. Cell 44, 419-428 (1986).
53. Hasty, P., Rivera-Perez, J. & Bradley, A. The length of homology required for gene targeting in embryonic stem cells. Mol Cell Biol 11, 5586-5591 (1991).
54. Wu, S., Ying, G.X., Wu, Q. & Capecchi, M.R. A protocol for constructing gene targeting vectors: generating knockout mice for the cadherin family and beyond. Nat Protoc 3, 1056-1076 (2008).

55. Oliveira, T.Y. et al. Translocation capture sequencing: a method for high throughput mapping of chromosomal rearrangements. J Immunol Methods 375, 176-181 (2012).
56. Tremblay et al., Transcription Activator-Like Effector Proteins Induce the Expression of the Frataxin Gene; Human Gene Therapy. August 2012, 23(8): 883-890.
57. Shalek et al. Nanowire-mediated delivery enables functional interrogation of primary immune cells: application to the analysis of chronic lymphocytic leukemia. Nano Letters, 2012, Dec 12;12(12):6498-504.
58. Pardridge et al. Preparation of Trojan horse liposomes (THLs) for gene transfer across the blood-brain barrier; Cold Spring Harb Protoc; 2010; Apr; 2010 (4)
59. Plosker GL et al. Fluvastatin: a review of its pharmacology and use in the management of hypercholesterolaemia; Drugs 1996, 51(3):433-459).
60. Trapani et al. Potential role of nonstatin cholesterol lowering agents; IUBMB Life, Volumen 63, Tema 11, páginas 964-971, November 2011
61. Birch AM et al. DGAT1 inhibitors as anti-obesity and anti-diabetic agents; Current Opinion in Drug Discovery & Development, 2010, 13(4):489-496
62. Fuchs et al. Killing of leukemic cells with a BCR/ABL fusion gene by RNA interference (RNAi), Oncogene 2002, 21(37):5716-5724.
63. McManaman JL et al. Perilipin-2 Null Mice are Protected Against Diet-Induced Obesity, Adipose Inflammation and Fatty Liver Disease; The Journal of Lipid Research, jlr.M035063. First Published on February 12, 2013.
64. Tang J et al. Inhibition of SREBP by a Small Molecule, Betulin, Improves Hyperlipidemia and Insulin Resistance and Reduces Atherosclerotic Plaques; Cell Metabolism, Volumen 13, Issue 1, 44-56, 5 January 2011.
65. Dumitrache et al. Trex2 enables spontaneous sister chromatid exchanges without facilitating DNA double-strand break repair; Genetics. 2011 August; 188(4): 787-797

REIVINDICACIONES

1. Un método implementado por computadora para seleccionar un complejo CRISPR para dirigir y/o escindir una secuencia de ácido nucleico diana candidata dentro de una célula, que comprende los pasos de:

(a) determinar la cantidad, la ubicación y la naturaleza de los desapareamientos de la secuencia guía del o de los complejos CRISPR potenciales y la secuencia de ácido nucleico diana candidata,
(b) determinar la contribución de cada cantidad, la ubicación y la naturaleza del o de los desapareamientos en la energía libre de hibridación de la unión entre la secuencia de ácido nucleico diana y la secuencia guía del o de los complejos CRISPR potenciales a partir de un conjunto de datos de capacitación,
(c) basándose en el análisis de contribución del paso (b), predecir la escisión en la o las ubicaciones del o de los desapareamientos de la secuencia del ácido nucleico diana por el o los complejos CRISPR potenciales, y
(d) seleccionar el complejo CRISPR del o de los complejos CRISPR potenciales en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no se produzca en las ubicaciones del o de los desapareamientos por el complejo CRISPR

en el que el paso (b) se realiza mediante
la definición de un modelo termodinámico que tiene un conjunto de pesos que vinculan la energía libre eficaz de hibridación Z a las energías libres locales G;
la definición de un conjunto de capacitación de pares de secuencia de ARN guía/ADN diana;
el ingreso de valores conocidos de energías libres locales G para cada par de secuencias de ARN guía/ADN diana en el conjunto de capacitación;
el cálculo de un valor de la energía libre eficaz de hibridación Z para cada par de secuencias de ARN guía/ADN diana en el conjunto de capacitación;
la determinación de los pesos utilizando un algoritmo de aprendizaje automático y
la emisión de los pesos mediante los cuales se pueden usar los pesos para estimar la energía libre de hibridación para cualquier secuencia.

2. El método de la reivindicación 1, en el que la secuencia diana candidata es una secuencia de ADN, y el o los desapareamientos son de ARN del o de los complejos CRISPR potenciales y el ADN.

3. El método de la reivindicación 1 o la reivindicación 2, en el que el paso (b) se realiza mediante
la determinación de las energías libres locales conocidas, ΔGij(k), entre cada secuencia de ARN guía i y la secuencia de ácido nucleico de ADN diana j en la posición k,
el cálculo de los valores de la energía libre eficaz Z_{ij} que utiliza la relación $p_{ij} \propto e^{-\beta Z_{ij}}$, donde p_{ij} es la frecuencia de corte medida mediante la secuencia de ARN guía i en la secuencia de ácido nucleico de ADN diana j en el conjunto de capacitación y β es una constante de proporcionalidad positiva,
la determinación de los pesos que son pesos dependientes de la posición α ajustando el valor conocido de ΔGij(k) y el valor calculado de Z_{ij} a través de cada par de secuencia de ARN guía/ADN diana en el conjunto de capacitación en la suma a través de todas las N bases de la secuencia guía

$$Z_{ij} = \sum_{k=1}^{N} \alpha_k \Delta G_{ij}(k)$$

mediante la escritura de la ecuación anterior en forma de matriz $\vec{Z} = G\vec{\alpha}$ y en donde, el paso (c) se realiza mediante la estimación de la energía libre eficaz Z_{est} utilizando los pesos dependientes de la posición determinada en la ecuación

$$\overrightarrow{Z_{est}} = G\vec{\alpha}$$

y la determinación de las frecuencias de corte espaciador-diana estimadas $p_{est} \propto e^{-\beta Z_{est}}$, para así predecir la escisión.

4. El método de una cualquiera de las reivindicaciones 1 a 3, que comprende además la normalización de los valores calculados de la energía libre eficaz de hibridación Z para cada par de secuencias de ARN guía/ADN diana en el conjunto de capacitación.

5. El método de una cualquiera de las reivindicaciones 1 a 4, que comprende además filtrar el valor calculado de la energía libre eficaz de hibridación Z para cada par de secuencias de ARN de guía/ADN diana en el conjunto de capacitación que tiene una profundidad de secuenciación que está por debajo de una profundidad de secuenciación mínima.

6. Un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, hacen que dichos uno o más procesadores implementen el método según lo establecido en una cualquiera de las reivindicaciones 1 a 5.

7. Un sistema informático para seleccionar un complejo CRISPR para dirigir y/o escindir una secuencia de ácido nucleico diana candidata dentro de una célula, comprendiendo el sistema:

a. una unidad de memoria configurada para recibir y/o almacenar información de secuencia de la secuencia de ácido nucleico diana candidata; y
b. uno o más procesadores solos o en combinación, programados para

(a) determinar la cantidad, la ubicación y la naturaleza del o de los desapareamientos del o de los complejos CRISPR potenciales y la secuencia de ácido nucleico diana candidata,
(b) determinar la contribución del o de los desapareamientos en función de la cantidad y la ubicación del o de los desapareamientos,
(c) basándose en el análisis de contribución del paso (b), predecir la escisión en la o las ubicaciones del o de los desapareamientos y
(d) seleccionar el complejo CRISPR del o de los complejos CRISPR potenciales en función de si la predicción del paso (c) indica que es más probable que se produzca la escisión a que no se produzca en la o las ubicaciones del o de los desapareamientos por el complejo CRISPR

en donde el paso (b) se realiza mediante
la definición de un modelo termodinámico que tiene un conjunto de pesos que vinculan la energía libre eficaz de hibridación Z a las energías libres locales G;
la definición de un conjunto de capacitación de pares de secuencia de ARN guía/ADN diana;
el ingreso de valores conocidos de energías libres locales G para cada par de secuencias de ARN guía/ADN diana en el conjunto de capacitación;
el cálculo de un valor de la energía libre eficaz de hibridación Z para cada par de secuencias de ARN guía/ADN diana en el conjunto de capacitación;
la determinación de los pesos utilizando un algoritmo de aprendizaje automático y
la emisión de los pesos mediante los cuales se pueden usar los pesos para estimar la energía libre de hibridación para cualquier secuencia.

8. Un método implementado por ordenador para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, mediante el cual la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende:

a) determinar la frecuencia de corte promedio de los desapareamientos ARN guía/diana en una posición particular para una Cas particular a partir de un conjunto de datos de capacitación en cuanto a esa Cas, y
b) determinar la frecuencia de corte promedio de un tipo de desapareamiento particular para la Cas particular a partir del conjunto de datos de capacitación,

para obtener así una clasificación, que permite la identificación de una o más secuencias diana únicas.

9. El método de la reivindicación 8 que comprende

c) multiplicar la frecuencia de corte promedio en una posición particular por la frecuencia de corte promedio de un tipo de desapareamiento particular para obtener un primer producto,
d) repetir los pasos a) a c) para obtener un segundo y más productos para cualquier posición o posiciones particulares adicionales de desapareamientos y desapareamientos particulares y multiplicar esos segundos y otros productos por el primer producto, para un producto final, y omitir este paso si no hay desapareamiento, en ninguna posición o si solo hay un desapareamiento particular en una posición particular, y
e) multiplicar el producto final por el resultado de dividir la distancia mínima entre desapareamientos consecutivos por la distancia, en pb, entre la primera y la última bases de la secuencia diana y omitir este paso si no hay desapareamientos en ninguna posición o si solo hay un desapareamiento particular en una posición particular, para obtener así la clasificación que permite la identificación de una o más secuencias diana únicas.

10. El método de la reivindicación 8 o la reivindicación 9, que comprende crear el conjunto de datos de capacitación con respecto a una Cas particular, antes de realizar el paso (a).

11. El método de una cualquiera de las reivindicaciones 2 a 5 y 8 a 10, en el que la distancia, en pb, entre la primera y la última bases de la secuencia diana es 18.

12. El método de una cualquiera de las reivindicaciones 8 a 11, en el que la frecuencia de corte promedio p_{est} se determina a partir de $p_{est} \propto e^{-\beta Z_{est}}$, donde β es una constante positiva de proporcionalidad y Z_{est} es la energía libre eficaz determinada utilizando $\overrightarrow{Z_{est}} = G\vec{\alpha}$ donde G es la energía libre local y α son pesos dependientes de la posición determinados utilizando el conjunto de capacitación.

13. Un medio legible por ordenador que comprende códigos que, al ser ejecutados por uno o más procesadores, implementa un método para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota,

mediante el cual la secuencia diana es susceptible de ser reconocida por un sistema CRISPR-Cas, en donde el método comprende un método de acuerdo con una cualquiera de las reivindicaciones 8 a 12.

14. Un sistema informático para identificar una o más secuencias diana únicas en un genoma de un organismo eucariota, comprendiendo el sistema:

I. una unidad de memoria configurada para recibir y/o almacenar información de secuencia del genoma; y
II. uno o más procesadores solos o en combinación, programados para realizar un método de acuerdo con una cualquiera de las reivindicaciones 8 a 12.

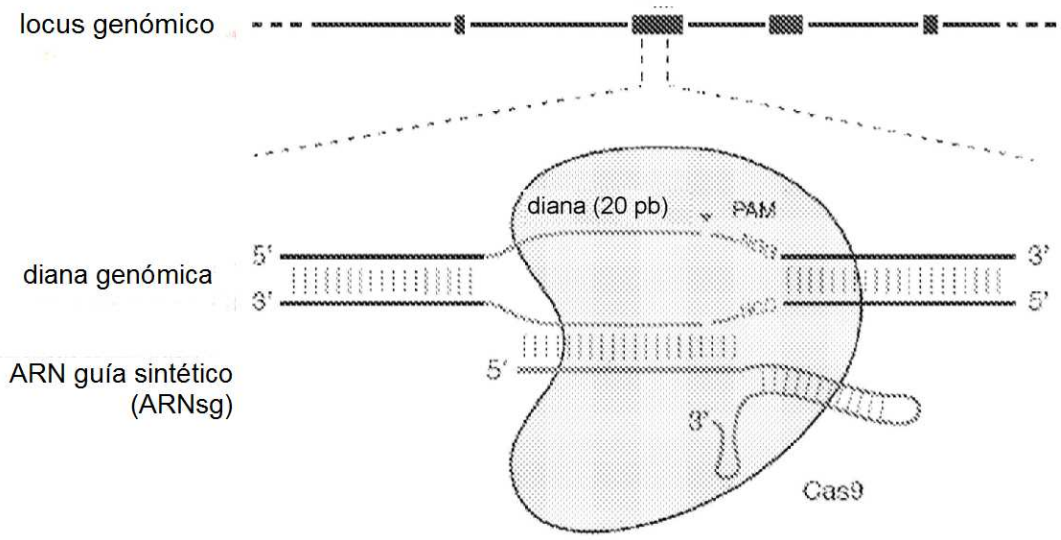

locus genómico
diana (20 pb)
PAM
diana genómica
5'
3'
3'
5'
ARN guía sintético
(ARNsg)
5'
3'
Cas9

FIG. 1

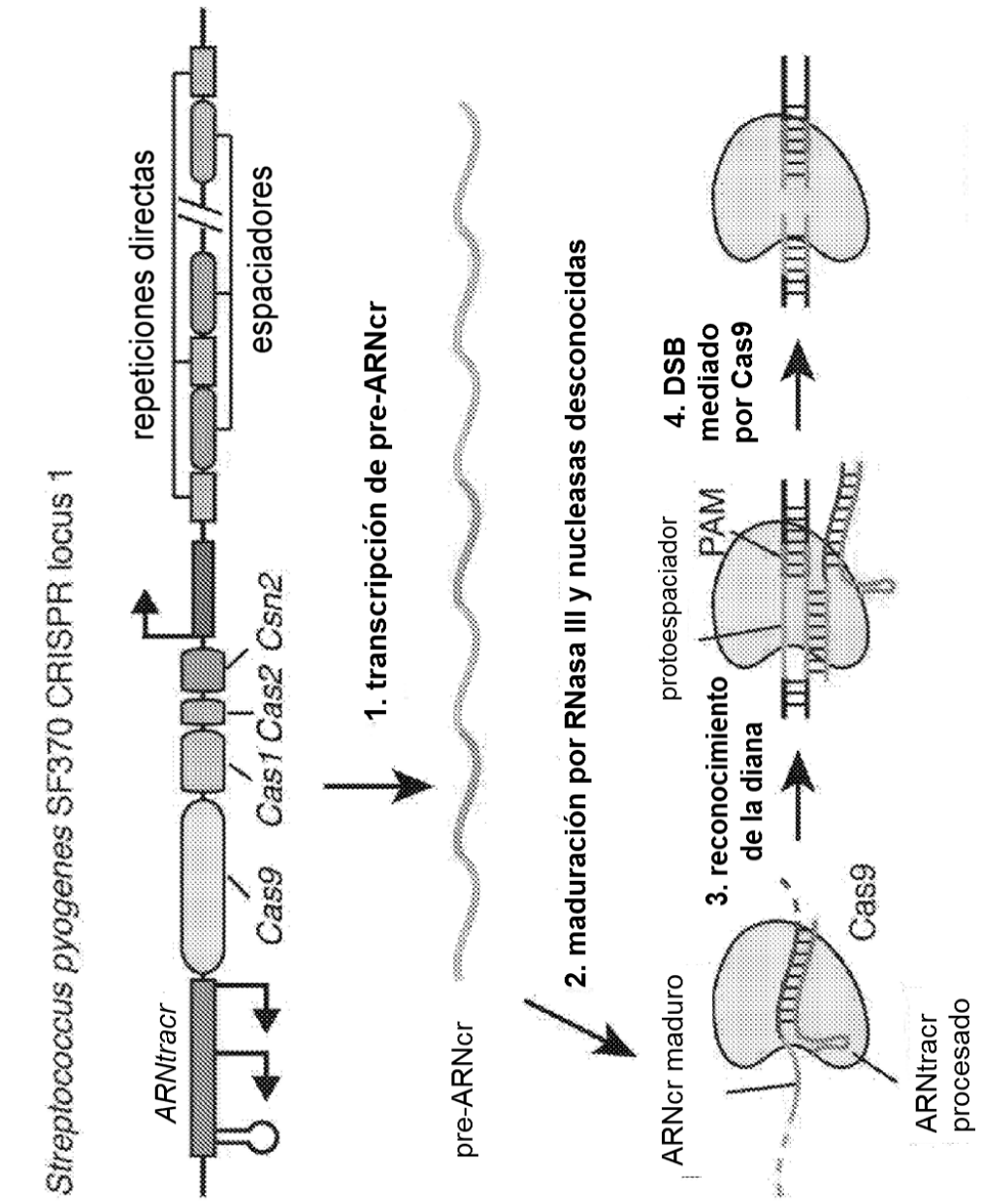

A
Streptococcus pyogenes SF370 CRISPR locus 1
repeticiones directas
espaciadores
ARNtracr
Cas9
Cas1
Cas2
Csn2
1. transcripción de pre-ARNcr
pre-ARNcr
2. maduración por RNasa III y nucleasas desconocidas
ARNcr maduro
protoespaciador
PAM
3. reconocimiento de la diana
4. DSB mediado por Cas9
Cas9
ARNtracr procesado

FIG. 2A

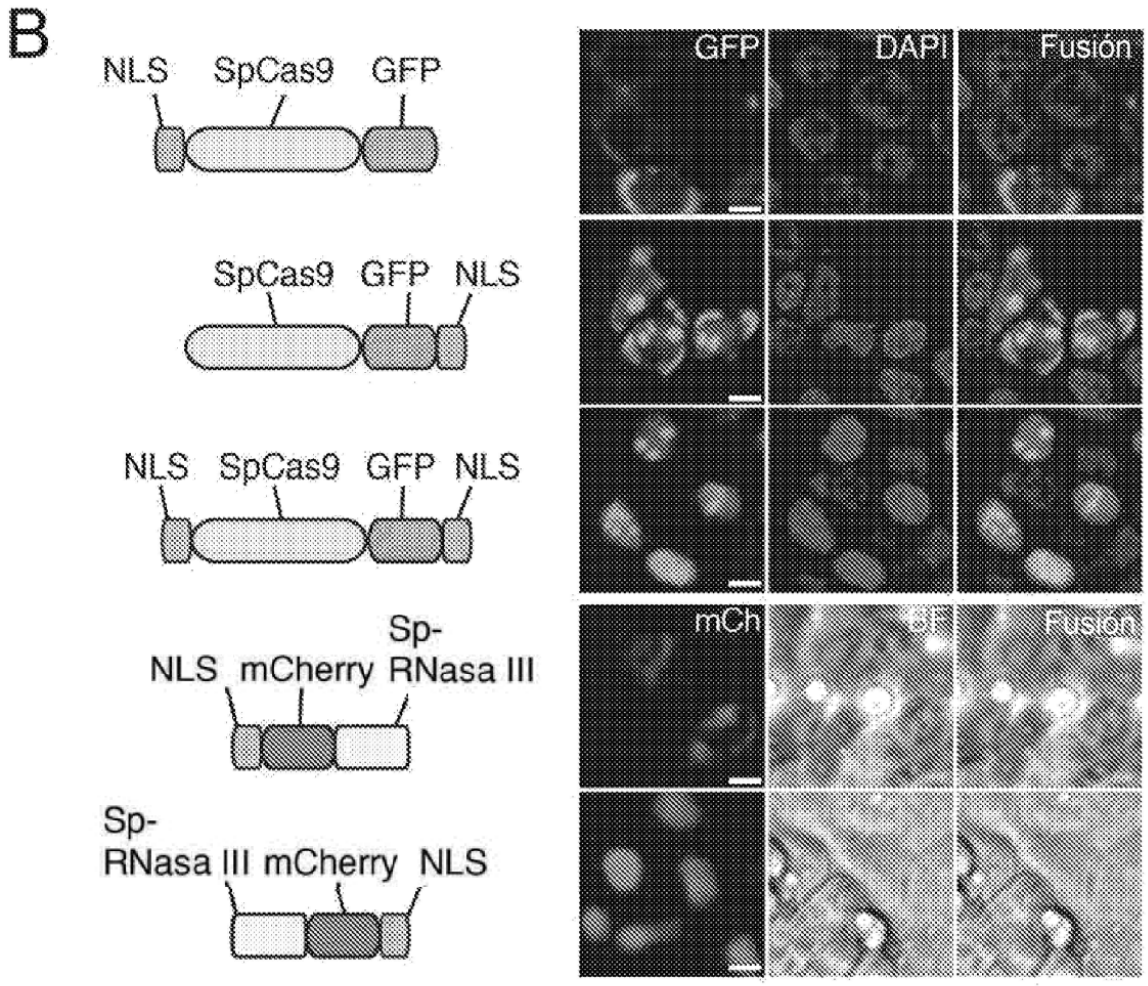

B
NLS SpCas9 GFP
SpCas9 GFP NLS
NLS SpCas9 GFP NLS
Sp-
NLS mCherry RNasa III
Sp-
RNasa III mCherry NLS
GFP
DAPI
Fusion
mCh
BF
Fusion

FIG. 2B

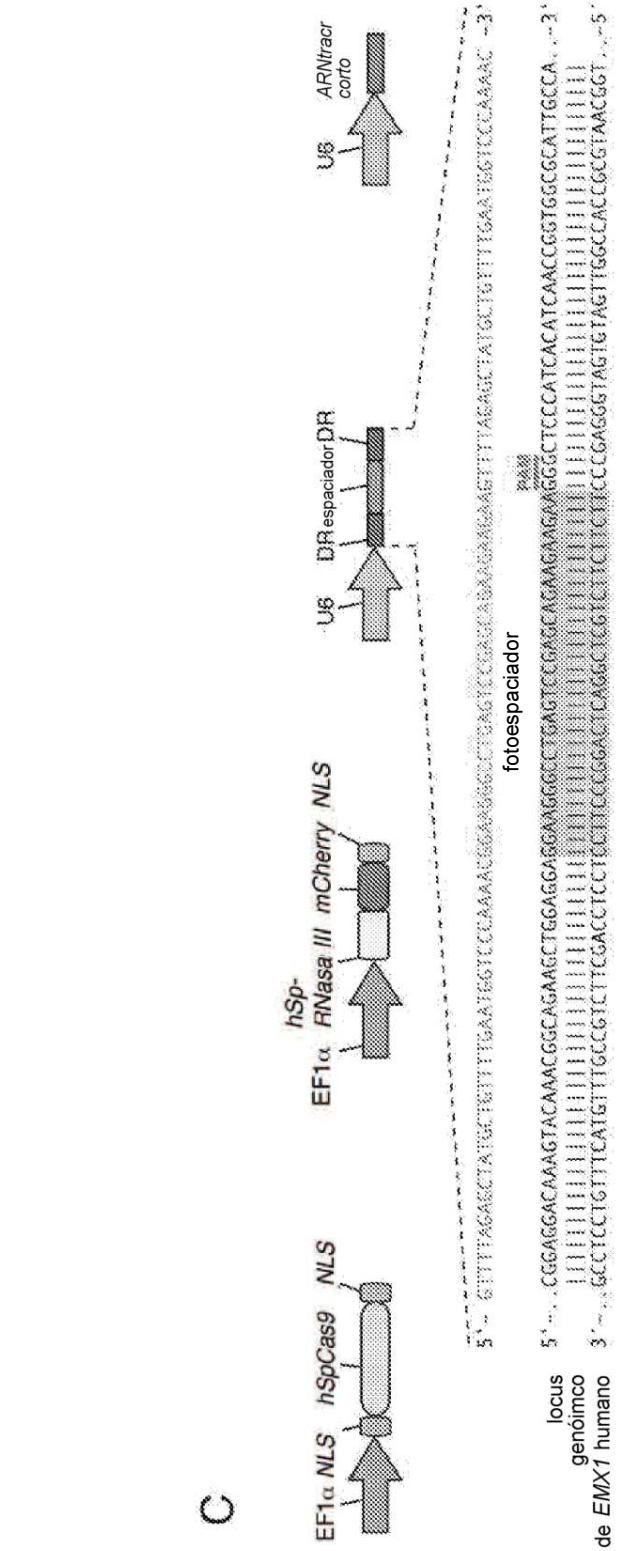

C
EF1α NLS hSpCas9 NLS
EF1α hSp-RNasa III mCherry NLS
U6 DR espaciador DR
U6 ARNtracr corto
fotoespaciador
PAM
locus genóimco de EMX1 humano
5'- ...CGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCA...-3'
3'- ...GCCTCCTGTTTCATGTTTGCCGTCTTCGACCTCCTCCCGGACTCAGGCTCGTCTTCTTCTCCCGAGGGTAGTGTAGTTGGCCACCGCGTAACGGT...-5'

FIG. 2C

D

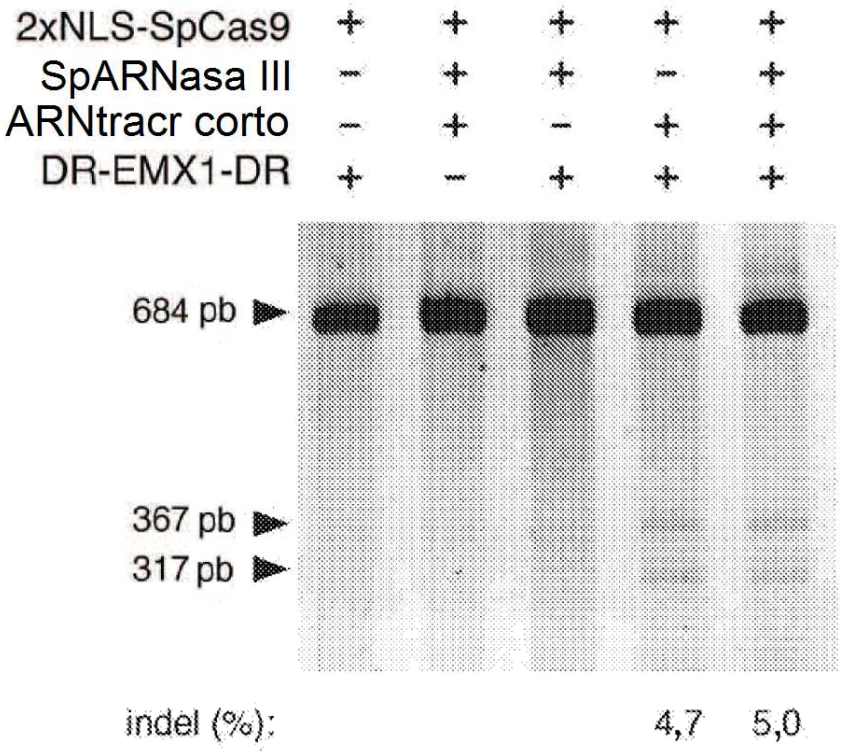

2xNLS-SpCas9 + + + + +
SpARNasa III − + + − +
ARNtracr corto − + − + +
DR-EMX1-DR + − + + +
684 pb
367 pb
317 pb
indel (%): 4,7 5,0

FIG. 2D

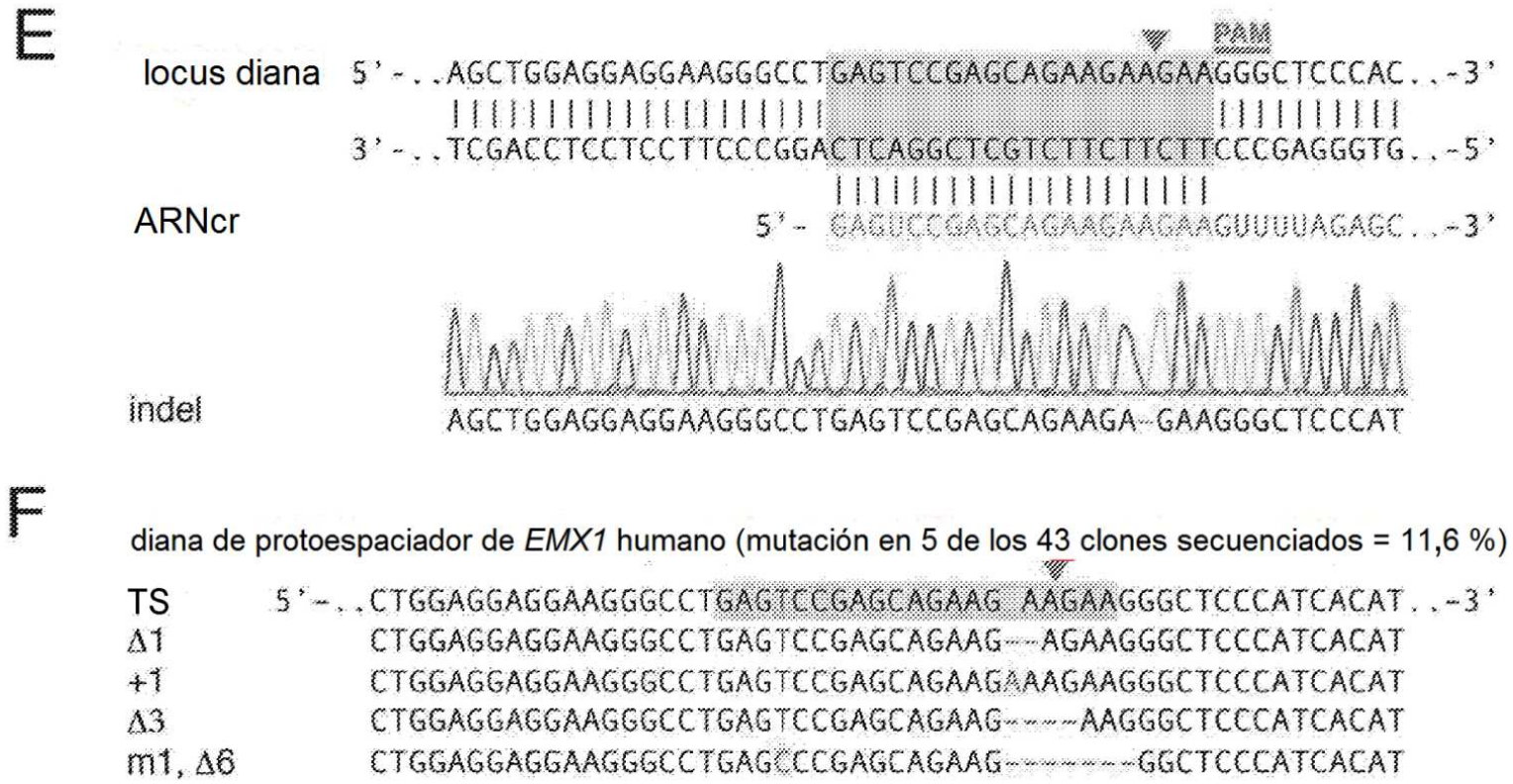

E
PAM
locus diana 5'-..AGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCAC..-3'
3'-..TCGACCTCCTCCTTCCCGGACTCAGGCTCGTCTTCTTCTTCCCGAGGGTG..-5'
ARNcr 5'- GAGUCCGAGCAGAAGAAGAAGUUUUAGAGC..-3'
indel
AGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGA-GAAGGGCTCCCAT
F
diana de protoespaciador de EMX1 humano (mutación en 5 de los 43 clones secuenciados = 11,6 %)
TS 5'-..CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG AAGAAGGGCTCCCATCACAT..-3'
Δ1 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG--AGAAGGGCTCCCATCACAT
+1 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAAGAAGGGCTCCCATCACAT
Δ3 CTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAG----AAGGGCTCCCATCACAT
m1, Δ6 CTGGAGGAGGAAGGGCCTGAGCCCGAGCAGAAG-------GGCTCCCATCACAT

FIG. 2E-F

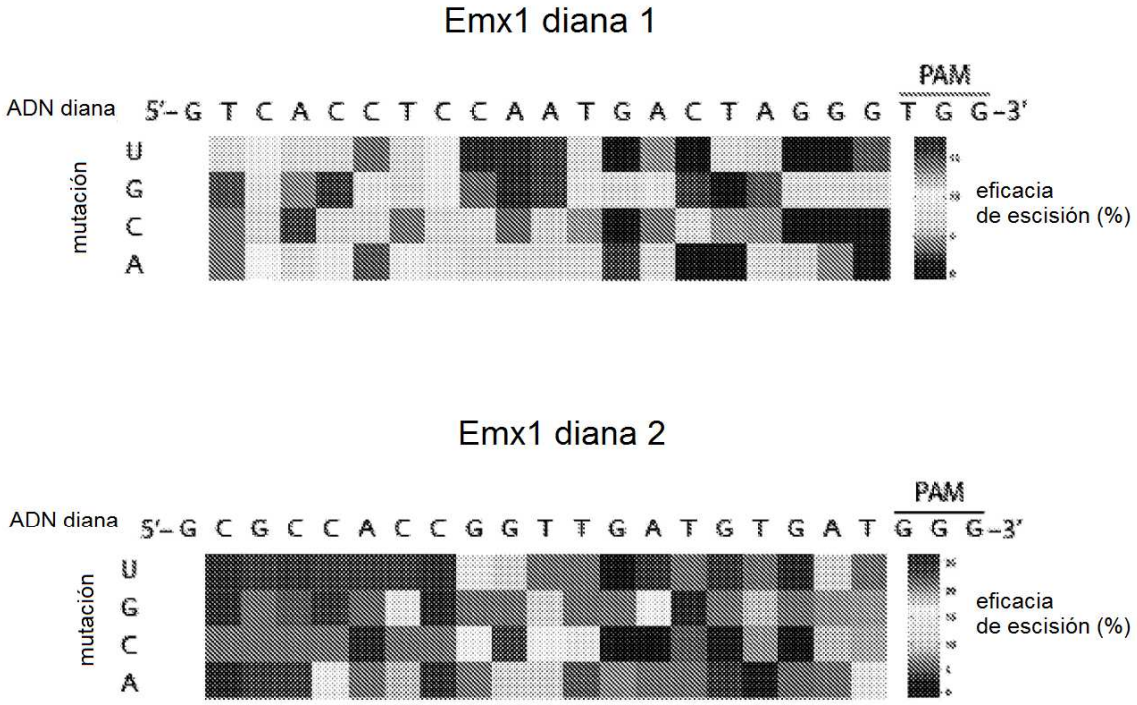

Emx1 diana 1
PAM
ADN diana 5'-G T C A C C T C C A A T G A C T A G G G T G G-3'
mutación
U
G
C
A
eficacia
de escisión (%)
Emx1 diana 2
PAM
ADN diana 5'-G C G C C A C C G G T T G A T G T G A T G G G-3'
mutación
U
G
C
A
eficacia
de escisión (%)

FIG. 3

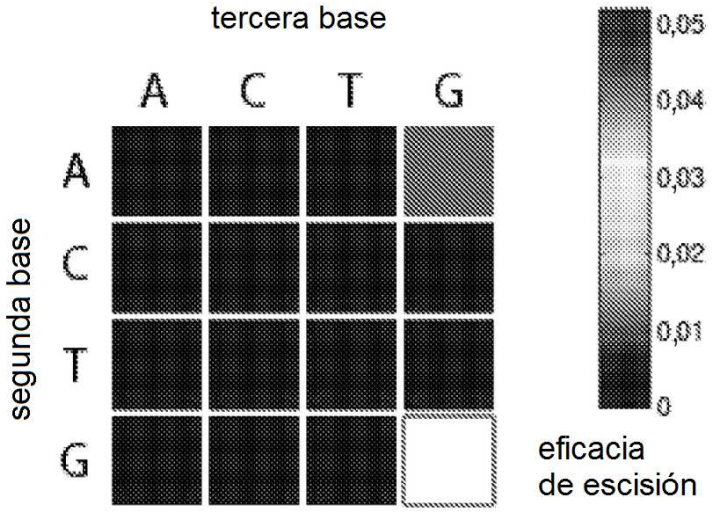
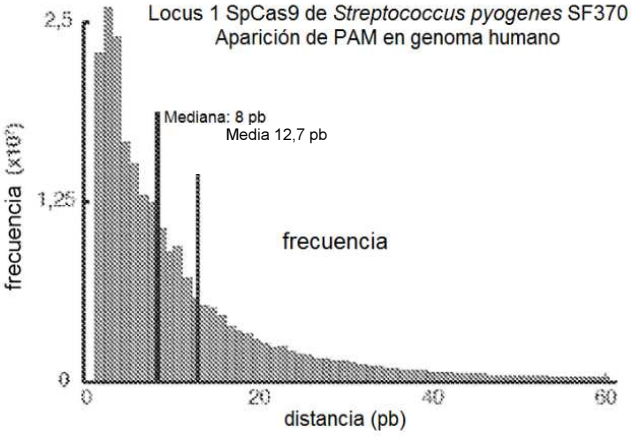
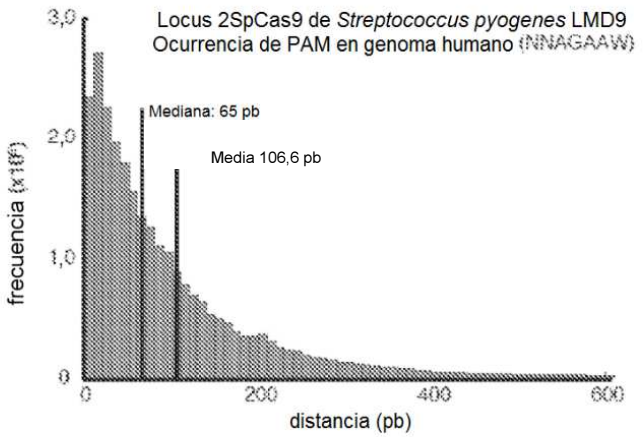

tercera base
A
C
T
G
segunda base
A
C
T
G
0,05
0,04
0,03
0,02
0,01
0
eficacia
de escisión

FIG. 4

A



B



C

	NGG		NNAGAAW	
Cr	mediana	media	mediana	media
1	7	12,8	67	115,8
2	8	12,7	64	100,8
3	8	13,0	63	98,5
4	9	14,0	61	94,5
5	8	13,1	63	97,9
6	8	13,1	63	98,5
7	8	12,4	64	102,9
8	8	12,8	64	100,9
9	7	13,9	65	120,5
10	7	12,1	66	107,0
11	7	12,0	65	105,8
12	8	12,4	65	103,5
13	8	13,6	62	94,6
14	8	12,0	65	101,5
15	7	11,5	68	107,7
16	7	11,7	74	136,8
17	6	10,3	76	127,9
18	8	13,4	63	101,8
19	6	9,4	82	145,4
20	7	11,1	72	121,8
21	7	13,4	64	111,4
22	6	9,2	85	140,3
X	8	13,2	63	99,0
Y	8	29,2	62	223,7

FIG. 5

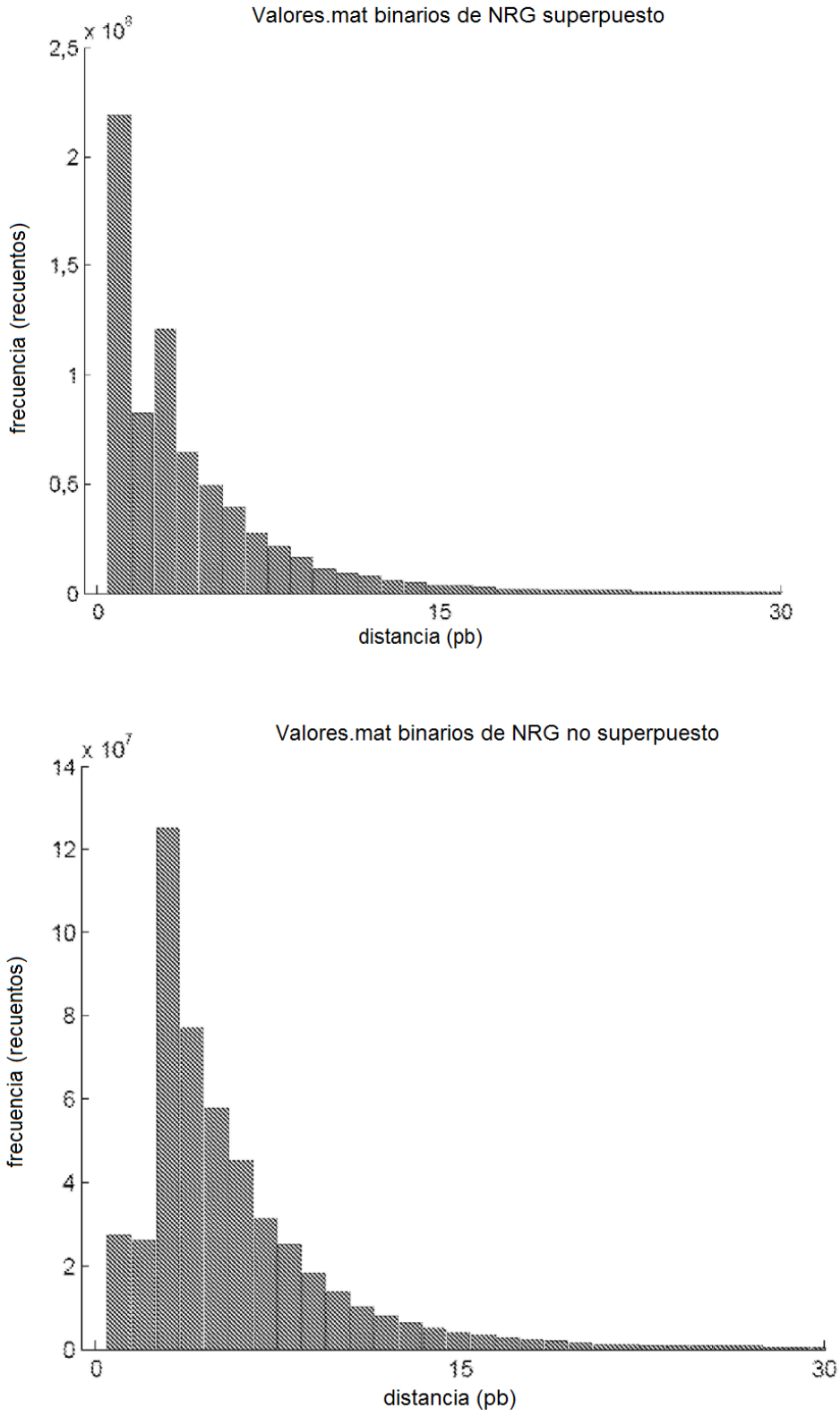

Valores.mat binarios de NRG superpuesto
x 10^3
2,5
2
1,5
1
0,5
0
frecuencia (recuentos)
0
15
30
distancia (pb)
Valores.mat binarios de NRG no superpuesto
x 10^7
14
12
10
8
6
4
2
0
frecuencia (recuentos)
0
15
30
distancia (pb)

FIG. 6A

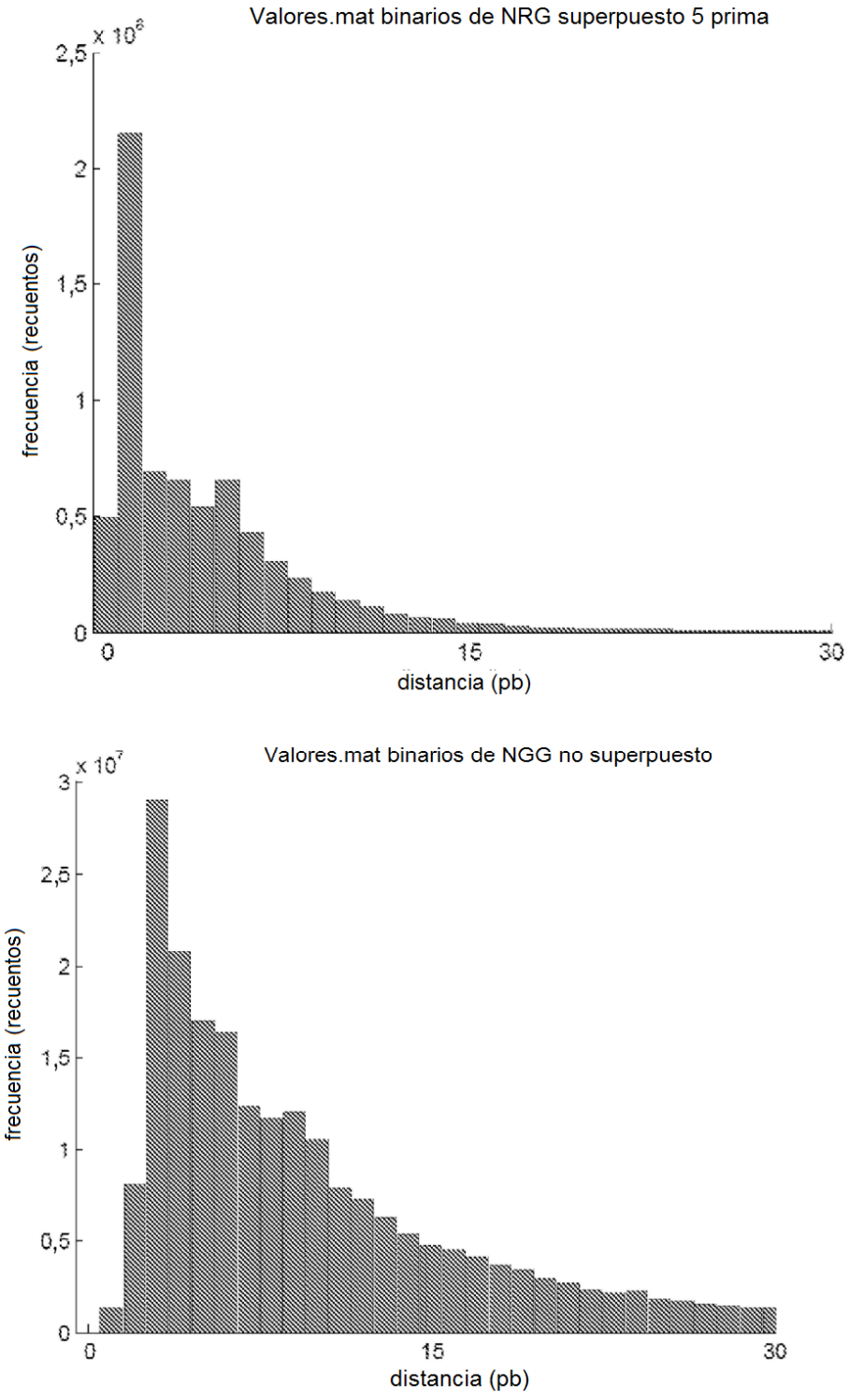
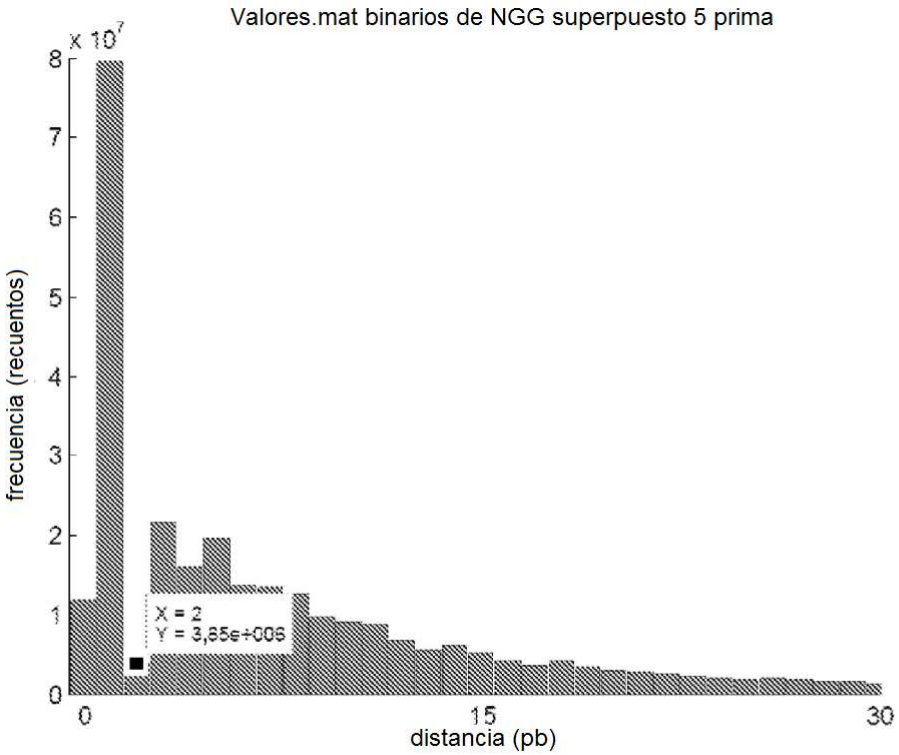

Valores.mat binarios de NRG superpuesto 5 prima
x 10^6
2,5
2
1,5
1
0,5
0
frecuencia (recuentos)
0
15
30
distancia (pb)
Valores.mat binarios de NGG no superpuesto
x 10^7
3
2,5
2
1,5
1
0,5
0
frecuencia (recuentos)
0
15
30
distancia (pb)

FIG. 6B


Valores.mat binarios de NGG superpuesto 5 prima
x 10^7
8
7
6
5
4
3
2
1
0
frecuencia (recuentos)
X = 2
Y = 3,85e+006
0
15
30
distancia (pb)

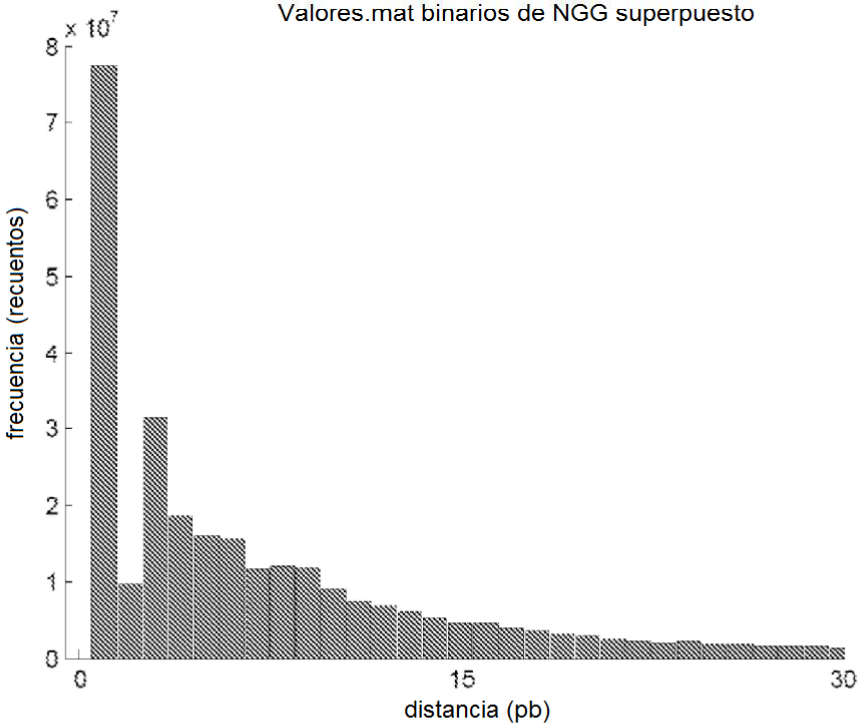

Valores.mat binarios de NGG superpuesto
x 10^7
8
7
6
5
4
3
2
1
0
frecuencia (recuentos)
0
15
30
distancia (pb)

FIG. 6C


proteína hipotética BBta_3952 [Bradyrhizobium sp. BTAi1]
proteína asociada a CRISPR, familia Csn1 [Candidatus Puniceispirillum marinum IMCC1322]
proteína asociada a CRISPR proteína de la familia Csn1 [Elusimicrobium minutum Pei191]
endonucleasa asociada a CRISPR, familia Csn1 [Sphaerochaeta globus str. Buddy]
proteína asociada a CRISPR similar a Csn1 [no cultivado Termite grupo 1 bacteria filotipo Rs-D17]
proteína asociada a CRISPR, familia Csn1 [Rhodospirillum rubrum ATCC 11170]
proteína asociada a CRISPR proteína de la familia Csn1 [Parvibaculum lavamentivorans DS-1]
endonucleasa asociada a CRISPR proteína de la familia Csn1 [Bacteroides sp. 20_3]
proteína hipotética BACDOR_03889 [Bacteroides coprophilus DSM 18228]
proteína hipotética BACCOPRO_03113 [Bacteroides coprosuis DSM 18777]
proteína hipotética [Fibrobacter succinogenes subsp. succinogenes S85]
proteína asociada a CRISPR, familia Csn1 [Prevotella bucca ATCC 33574]
proteína asociada a CRISPR, familia Csn1 [Prevotella melaninogenica ATCC 25845]
proteína hipotética conservada [Prevotella veroralis F0319]
proteína asociada a CRISPR, familia Csn1 [Prevotella bivia JCVIHMP022]
proteína asociada a CRISPR, familia Csn1 [Prevotella taffensis DSM 16823]
proteína asociada a CRISPR, familia Csn1 [Weeksella virosa DSM 16922]
proteína asociada a CRISPR, familia Csn1 [Flavobacterium columnare ATCC 49512]
proteína asociada a CRISPR, familia Csn1 [Sphingobacterium spiritivorum ATCC 33861]
proteína asociada a CRISPR, familia Csn1 [Mucilaginibacter paludis DSM 18603]
proteína asociada a CRISPR, familia Csn1 [Fluviicola taffensis DSM 16823]
proteína asociada a CRISPR, familia Csn1 [Capnocytophaga gingivalis ATCC 33624]
proteína asociada a CRISPR, familia Csn1 [Capnocytophaga ochracea DSM 7271]
proteína asociada a CRISPR, familia Csn1 [Capnocytophaga canimorsus Cc5]
proteína asociada a CRISPR, familia Csn1 [Bacteroidetes oral taxon 274 str. F0058]
proteína hipotética Ccan_14650 [Capnocytophaga ochracea DSM 7271]
proteína hipotética KAOT1_02362 [Kordia algicida OT-1]
proteína hipotética ZPR_4422 [Zunongwangia profunda SM-A87]
proteína asociada a CRISPR, familia Csn1 [Flavobacterium psychrophilum JIP02/86]
proteína de la familia Csn1 [Bacteroides cellulosilyticus DSM 14838]
proteína hipotética BACCELL_00696 [Parabacteroides johnsonii DSM 18315]
endonucleasa asociada a CRISPR, proteína hipotética PRABACTJOHN_00621 [Parabacteroides sp. D13]
proteína asociada a CRISPR, familia Csn1 [Bacteroides fragilis NCTC 9343]
proteína hipotética BF3954 [Bacteroides fragilis NCTC 9343]
proteína asociada a CRISPR, familia Csn1 [Prevotella oralis ATCC 33269]
proteína asociada a CRISPR, familia Csn1 [Prevotella timonensis CRIS 5C-B1]
proteína asociada a CRISPR, familia Csn1 [Prevotella buccalis ATCC 35310]
proteína asociada a CRISPR, familia Csn1 [Prevotella tannerae ATCC 51259]
proteína asociada a CRISPR, familia Csn1 [Prevotella ruminicola 23]
proteína asociada CRISPR supuesta (Cas) [Flavobacterium columnare ATCC 49512]
proteína hipotética HMPREF9140_01667 [Flavobacterium branchiophilum FL-15]
proteína asociada a CRISPR, familia Csn1 [Prevotella micans F0438]
proteína hipotética DIP0036 [Corynebacterium accolens ATCC 49726]
proteína hipotética conservada [Corynebacterium diphtheriae NCTC 13129]
exopolifosfatasa [Actinomyces coleocanis DSM 15436]
proteína asociada a CRISPR, familia Csn1 [Mobiluncus curtisii subsp. holmesii ATCC 35242]
endonucleasa de restricción [Bifidobacterium longum DJO10A]
endonucleasa asociada a CRISPR Csn1 [Bifidobacterium dentium Bd1]
proteína de dominio de endonucleasa HNH supuesta [Scardovia inopinata F0304]
nucleasa Csn1 asociada a CRISPR similar a McrA/HNH [Gardnerella vaginalis 5-1]
endonucleasa HNH [Acidothermus cellulolyticus 11B]
proteina hipotética Amuc_2010 [Akkermansia muciniphila ATCC BAA-835]

FIG. 7A


proteína hipotética Nham_4054 [Nitrobacter hamburgensis X14]
proteína asociada a CRISPR, familia Csn1 [Azospirillum sp. B510]
proteína hipotética SU517_3938 [Sphingomonas sp. SU17]
proteína hipotética conservada [Ralstonia syzygii R24]
endonucleasa asociada a CRISPR, familia Csn1 [gamma proteobacterium HdN1]
proteína asociada a CRISPR, proteína de la familia Csn1 [Nitrosomonas sp. AL212]
proteína asociada a CRISPR, familia Csn1 [Acidovorax avenae subsp. avenae ATCC 19860]
proteína asociada a CRISPR, familia Csn1 [Alicycliphilus denitrificans BC]
proteína asociada a CRISPR, familia Csn1 [Helicobacter mustelae 12198]
proteína de la familia Csn1 [Campylobacter canadensis MIT 98-5491]
proteína asociada a CRISPR, familia Csn1 [Gluconacetobacter diazotrophicus PAl 5]
proteína asociada a CRISPR, familia Csn1 [Rhodopseudomonas palustris BisB18]
proteína asociada a CRISPR, familia Csn1 [Methylocystis sp. ATCC 49242]
proteína asociada a CRISPR, familia Csn1 [Clostridium cellulolyticum H10]
proteína asociada a CRISPR, familia Csn1 [Lactobacillus coryniformis subsp. torquens KCTC 3535]
proteína asociada a CRISPR, familia Csn1 [Roseburia intestinalis L1-82]
proteína asociada a CRISPR, familia Csn1 [Aminomonas paucivorans DSM 12260]
proteína asociada a CRISPR, familia Csn1 [Simonsiella muelleri ATCC 29453]
proteína asociada a CRISPR, familia Csn1 [Neisseria flavescens SK114]
proteína asociada a CRISPR, familia Csn1 [Neisseria cinerea ATCC 14685]
proteína asociada a CRISPR, familia Csn1 [Neisseria lactamica 020-06]
proteína asociada a CRISPR, familia Csn1 [Neisseria meningitidis 053442]
proteína asociada a CRISPR, familia Csn1 [Neisseria meningitidis Z2491]
proteína asociada a CRISPR, familia Csn1 [Neisseria meningitidis alpha14]
proteína asociada a CRISPR, familia Csn1 [Kingella kingae ATCC 23330]
proteína hipotética NLA_17660 [Neisseria bacilliformis ATCC BAA-1200]
proteína hipotética NMCC_0397 [Neisseria meningitidis 8013]
proteína hipotética NMA0631 [Neisseria meningitidis alpha14]
proteína asociada a CRISPR supuesta [Neisseria bacilliformis ATCC BAA-1200]
proteína hipotética HMPREF9123_0020 [Neisseria de la familia Csn1 [Kingella kingae ATCC 23330]
endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Actinobacillus minor NM305]
proteína hipotética PM1127 [Pasteurella multocida subsp. multocida str. Pm70]
endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Actinobacillus pleuropneumoniae serovar 10 str. D13039]
proteína hipotética PARA_18570 [Haemophilus parainfluenzae T3T1]
proteína asociada a CRISPR, familia Csn1 [Actinobacillus succinogenes 130Z]
endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Ruminococcus albus 8]
proteína asociada a CRISPR, familia Csn1 [Nitratifractor salsuginis DSM 16511]
proteína hipotética MMOB0330 [Mycoplasma mobile 163K]

FIG. 7B

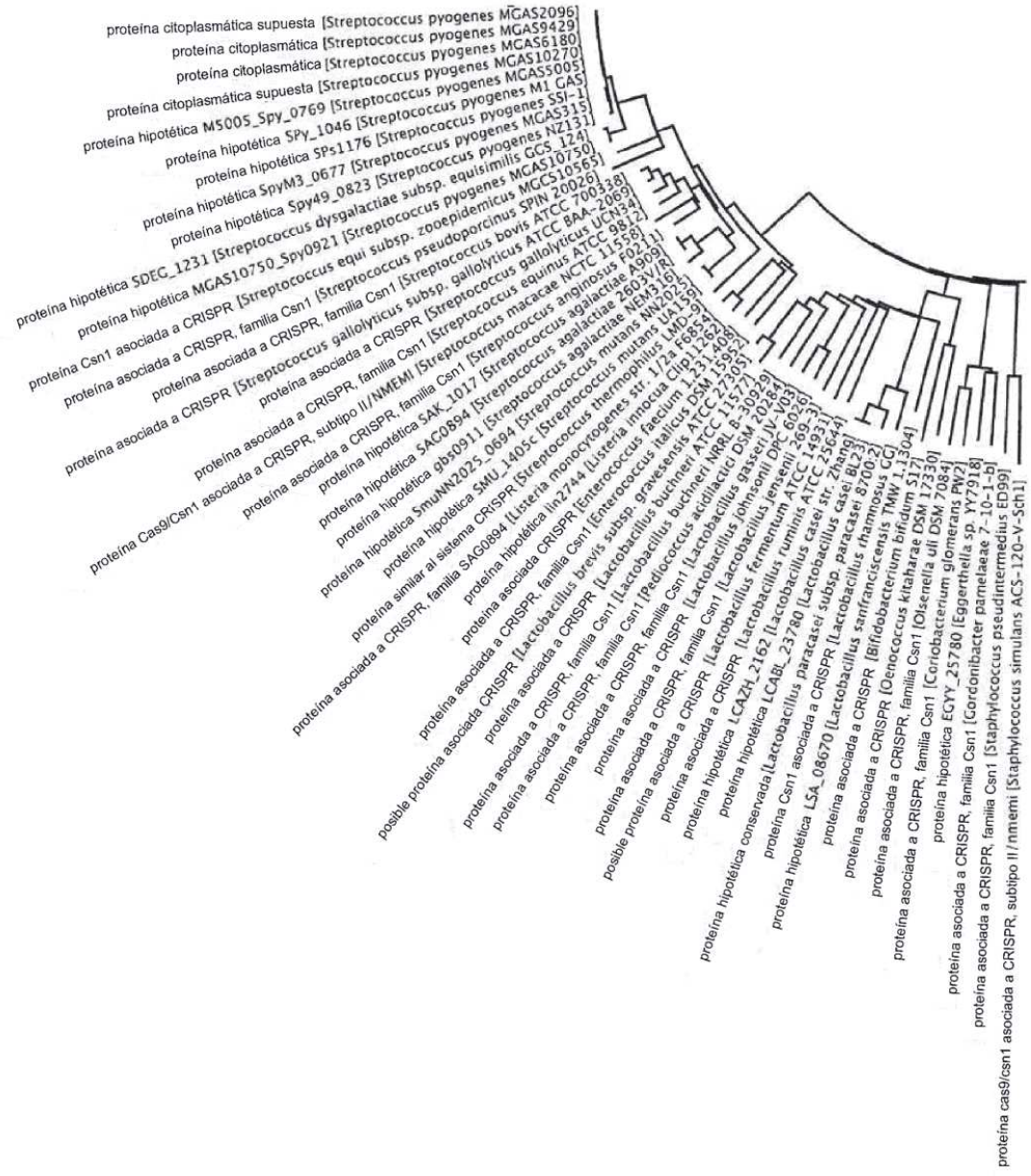

proteína citoplasmática supuesta [Streptococcus pyogenes MGAS2096]
proteína citoplasmática [Streptococcus pyogenes MGAS9429]
proteína citoplasmática [Streptococcus pyogenes MGAS6180]
proteína citoplasmática supuesta [Streptococcus pyogenes MGAS10270]
proteína hipotética M5005_Spy_0769 [Streptococcus pyogenes M1 GAS]
proteína hipotética SPy_1046 [Streptococcus pyogenes MGAS5005]
proteína hipotética SPs1176 [Streptococcus pyogenes SSI-1]
proteína hipotética SpyM3_0677 [Streptococcus pyogenes MGAS315]
proteína hipotética Spy49_0823 [Streptococcus pyogenes NZ131]
proteína hipotética SDEG_1231 [Streptococcus pyogenes MGAS10750]
proteína hipotética MGAS10750_Spy0921 [Streptococcus dysgalactiae subsp. equisimilis GGS_124]
proteína Csn1 asociada a CRISPR [Streptococcus equi subsp. zooepidemicus MGCS10565]
proteína asociada a CRISPR, familia Csn1 [Streptococcus pseudoporcinus SPIN 20026]
proteína asociada a CRISPR [Streptococcus gallolyticus subsp. gallolyticus ATCC 700338]
proteína asociada a CRISPR, familia Csn1 [Streptococcus bovis ATCC BAA-2069]
proteína Cas9/Csn1 asociada a CRISPR, subtipo II/NMEMI [Streptococcus equinus ATCC 9812]
proteína asociada a CRISPR, familia Csn1 [Streptococcus macacae NCTC 11558]
proteína hipotética SAK_1017 [Streptococcus agalactiae FO27-1]
proteína hipotética SAC0894 [Streptococcus agalactiae A909]
proteína hipotética gbs0911 [Streptococcus agalactiae 2603V/R]
proteína hipotética SmuNN2025_0694 [Streptococcus mutans NN2025]
proteína hipotética SMU_1405c [Streptococcus mutans UA159]
proteína similar al sistema CRISPR [Streptococcus thermophilus LMD-9]
proteína hipotética SAG0894 [Listeria monocytogenes str. 1/2a F6854]
proteína hipotética lin2744 [Listeria innocua Clip11262]
proteína asociada a CRISPR, familia Csn1 [Enterococcus faecium 1,231,408]
proteína asociada a CRISPR, familia Csn1 [Lactobacillus brevis subsp. gravesensis ATCC 27305]
proteína asociada a CRISPR, familia Csn1 [Pediococcus acidilactici DSM 20284]
proteína asociada a CRISPR, familia Csn1 [Lactobacillus buchneri NRRL B-30929]
proteína asociada a CRISPR, familia Csn1 [Lactobacillus johnsonii DPC 6026]
proteína asociada a CRISPR [Lactobacillus fermentum ATCC 14931]
proteína hipotética LCAZH_2162 [Lactobacillus jensenii 269-3]
proteína hipotética LCABL_23780 [Lactobacillus ruminis ATCC 25644]
proteína conservada [Lactobacillus paracasei subsp. paracasei 8700:2]
proteína Csn1 asociada a CRISPR [Lactobacillus casei str. Zhang]
proteína hipotética LSA_08670 [Lactobacillus sanfranciscensis TMW 1.1304]
proteína asociada a CRISPR [Bifidobacterium bifidum S17]
proteína asociada a CRISPR, familia Csn1 [Oenococcus kitaharae DSM 17330]
proteína hipotética EGYY_25780 [Coriobacterium glomerans PW2]
proteína asociada a CRISPR, familia Csn1 [Olsenella uli DSM 7084]
proteína asociada a CRISPR, familia Csn1 [Gordonibacter pamelaeae 7-10-1-b]
proteína cas9/csn1 asociada a CRISPR, subtipo II/nmemi [Staphylococcus pseudintermedius ED99]
proteína asociada a CRISPR, familia Csn1 [Staphylococcus simulans ACS-120-V-Sch1]

FIG. 7C

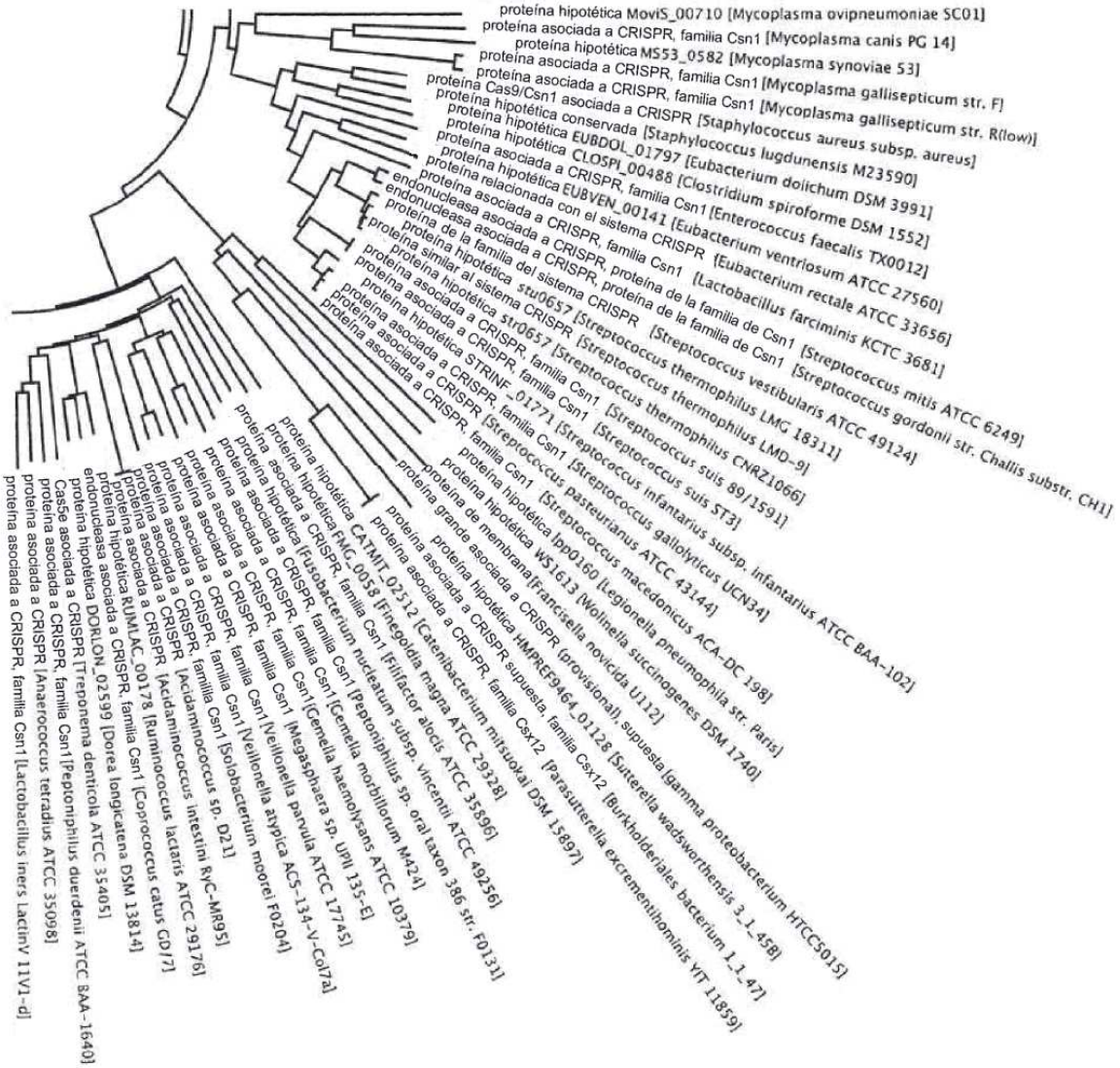

proteína hipotética MoviS_00710 [Mycoplasma ovipneumoniae SC01]
proteína asociada a CRISPR, familia Csn1 [Mycoplasma canis PG 14]
proteína hipotética MS53_0582 [Mycoplasma synoviae 53]
proteína asociada a CRISPR, familia Csn1 [Mycoplasma gallisepticum str. F]
proteína Cas9/Csn1 asociada a CRISPR [Mycoplasma gallisepticum str. R(low)]
proteína hipotética conservada [Staphylococcus aureus subsp. aureus]
proteína hipotética EUBDOL_01797 [Eubacterium dolichum DSM 3991]
proteína hipotética CLOSPI_00488 [Clostridium spiroforme DSM 1552]
proteína asociada a CRISPR, familia Csn1 [Enterococcus faecalis TX0012]
proteína hipotética EUBVEN_00141 [Eubacterium ventriosum ATCC 27560]
proteína relacionada con el sistema CRISPR [Eubacterium rectale ATCC 33656]
endonucleasa asociada a CRISPR, familia Csn1 [Lactobacillus farciminis KCTC 3681]
proteína asociada a CRISPR, proteína de la familia de Csn1 [Streptococcus mitis ATCC 6249]
proteína de la familia del sistema CRISPR [Streptococcus gordonii str. Challis substr. CH1]
proteína hipotética str0657 [Streptococcus thermophilus LMG 18311]
proteína similar al sistema CRISPR [Streptococcus thermophilus CNRZ1066]
proteína hipotética STRINF_01771 [Streptococcus thermophilus LMD-9]
proteína asociada a CRISPR, familia Csn1 [Streptococcus suis 89/1591]
proteína asociada a CRISPR [Streptococcus pasteurianus ATCC 43144]
proteína hipotética lpp0160 [Legionella pneumophila str. Paris]
proteína hipotética WS1613 [Wolinella succinogenes DSM 1740]
proteína de membrana grande asociada a CRISPR [Francisella novicida U112]
proteína asociada a CRISPR, familia Csx12 [Parasutterella excrementihominis YIT 11859]
proteína hipotética HMPREF9464_01128 [Sutterella wadsworthensis 3_1_45B]
proteína hipotética FMG_0058 [Fusobacterium nucleatum subsp. vincentii ATCC 49256]
proteína hipotética CATMIT_02512 [Catenibacterium mitsuokai DSM 15897]
proteína asociada a CRISPR, familia Csn1 [Peptoniphilus sp. oral taxon 386 str. F0131]
proteína asociada a CRISPR, familia Csn1 [Finegoldia magna ATCC 29328]
proteína asociada a CRISPR, familia Csn1 [Filifactor alocis ATCC 35896]
proteína asociada a CRISPR, familia Csn1 [Gemella haemolysans ATCC 10379]
proteína asociada a CRISPR, familia Csn1 [Gemella morbillorum M424]
proteína asociada a CRISPR, familia Csn1 [Megasphaera sp. UPII 135-E]
proteína asociada a CRISPR, familia Csn1 [Veillonella atypica ACS-134-V-Col7a]
proteína asociada a CRISPR, familia Csn1 [Veillonella parvula ATCC 17745]
proteína asociada a CRISPR, familia Csn1 [Solobacterium moorei F0204]
proteína asociada a CRISPR, familia Csn1 [Acidaminococcus sp. D21]
proteína asociada a CRISPR [Acidaminococcus intestini RyC-MR95]
proteína asociada a CRISPR [Ruminococcus lactaris ATCC 29176]
proteína hipotética RUMLAC_00178 [Ruminococcus torques L2-14]
proteína asociada a CRISPR, familia Csn1 [Coprococcus catus GD/7]
proteína asociada a CRISPR, familia Csn1 [Dorea longicatena DSM 13814]
endonucleasa asociada a CRISPR [Treponema denticola ATCC 35405]
proteína hipotética DORLON_02599 [Peptoniphilus duerdenii ATCC BAA-1640]
Cas5e asociada a CRISPR, familia Csn1 [Peptoniphilus sp. oral taxon 386 str. F0131]
proteína asociada a CRISPR [Anaerococcus tetradius ATCC 35098]
proteína asociada a CRISPR, familia Csn1 [Lactobacillus iners LactinV 11V1-d]

FIG. 7D

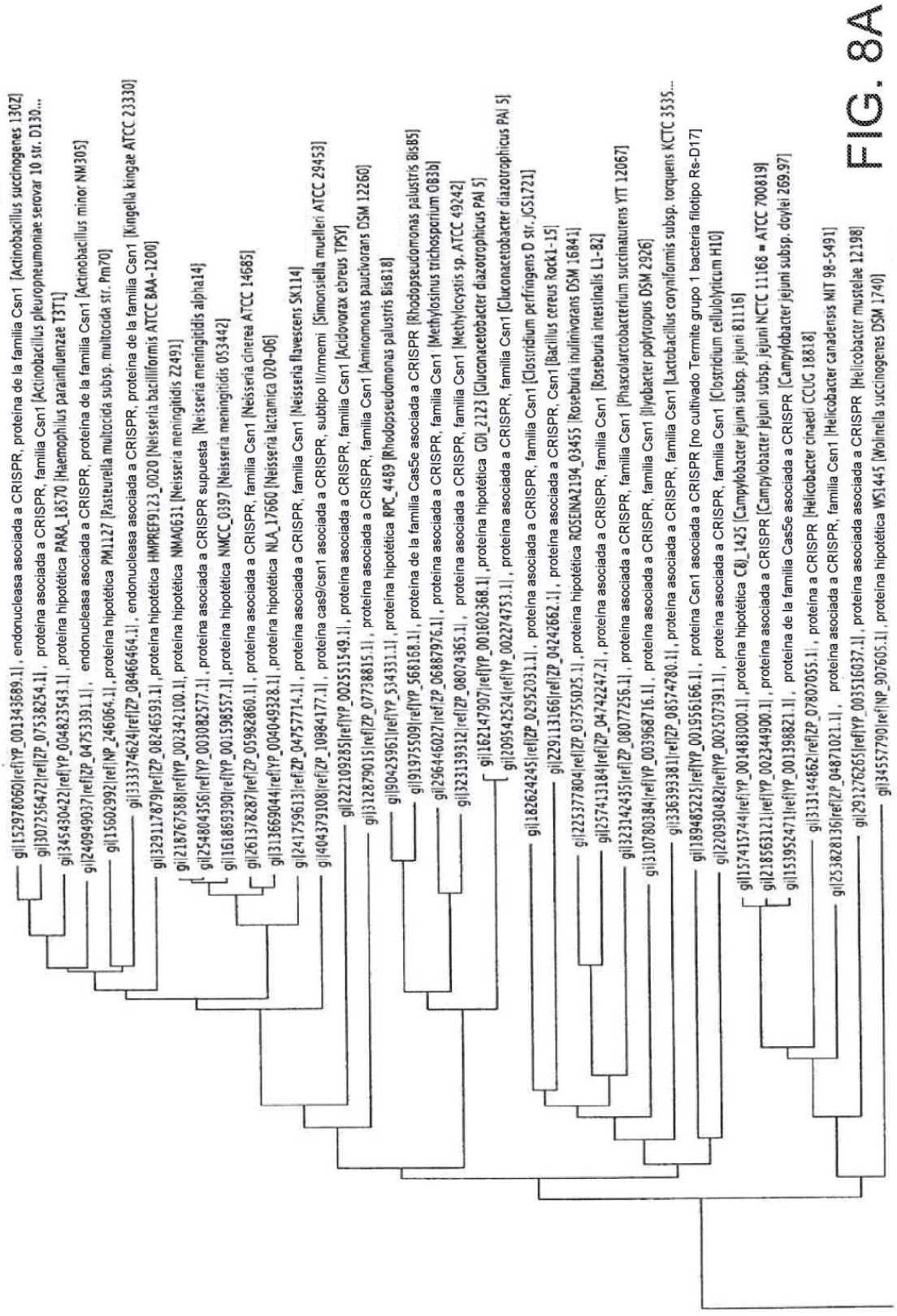

gi|152978060|ref|YP_001343689.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Actinobacillus succinogenes 130Z]
gi|307256472|ref|ZP_07538254.1|, proteína asociada a CRISPR, familia Csn1 [Actinobacillus pleuropneumoniae serovar 10 str. D130...
gi|345430422|ref|YP_004823543.1|, proteína hipotética PARA_18570 [Haemophilus parainfluenzae T3T1]
gi|240949037|ref|ZP_04753391.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Actinobacillus minor NM305]
gi|15602992|ref|NP_246064.1|, proteína hipotética PM1127 [Pasteurella multocida subsp. multocida str. Pm70]
gi|333374624|ref|ZP_08466464.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Kingella kingae ATCC 23330]
gi|329117879|ref|ZP_08246593.1|, proteína hipotética HMPREF9123_0020 [Neisseria bacilliformis ATCC BAA-1200]
gi|218767588|ref|YP_002342100.1|, proteína hipotética NMA0631 [Neisseria meningitidis Z2491]
gi|254804356|ref|YP_003082577.1|, proteína asociada a CRISPR supuesta [Neisseria meningitidis alpha14]
gi|161869390|ref|YP_001598557.1|, proteína hipotética NMCC_0397 [Neisseria meningitidis 053442]
gi|261378287|ref|ZP_05982860.1|, proteína asociada a CRISPR, familia Csn1 [Neisseria cinerea ATCC 14685]
gi|313669044|ref|YP_004049328.1|, proteína hipotética NLA_17660 [Neisseria lactamica 020-06]
gi|241759613|ref|ZP_04757714.1|, proteína asociada a CRISPR, familia Csn1 [Neisseria flavescens SK114]
gi|404379108|ref|ZP_10984177.1|, proteína cas9/csn1 asociada a CRISPR, subtipo II/nmemi [Simonsiella muelleri ATCC 29453]
gi|222109285|ref|YP_002551549.1|, proteína asociada a CRISPR, familia Csn1 [Acidovorax ebreus TPSY]
gi|312879015|ref|ZP_07738815.1|, proteína asociada a CRISPR, familia Csn1 [Aminomonas paucivorans DSM 12260]
gi|90425961|ref|YP_534331.1|, proteína hipotética RPC_4489 [Rhodopseudomonas palustris BisB18]
gi|91975509|ref|YP_568158.1|, proteína de la familia Cas5e asociada a CRISPR [Rhodopseudomonas palustris BisB5]
gi|296446027|ref|ZP_06887976.1|, proteína asociada a CRISPR, familia Csn1 [Methylosinus trichosporium OB3b]
gi|323139312|ref|ZP_08074365.1|, proteína asociada a CRISPR, familia Csn1 [Methylocystis sp. ATCC 49242]
gi|162147907|ref|YP_001602368.1|, proteína hipotética GDI_2123 [Gluconacetobacter diazotrophicus PAl 5]
gi|209542524|ref|YP_002274753.1|, proteína asociada a CRISPR, familia Csn1 [Gluconacetobacter diazotrophicus PAl 5]
gi|182624245|ref|ZP_02952031.1|, proteína asociada a CRISPR, familia Csn1 [Clostridium perfringens D str. JGS1721]
gi|229113166|ref|ZP_04242662.1|, proteína asociada a CRISPR, Csn1 [Bacillus cereus Rock1-15]
gi|225377804|ref|ZP_03755025.1|, proteína hipotética ROSEINA2194_03455 [Roseburia inulinivorans DSM 16841]
gi|257413184|ref|ZP_04742247.2|, proteína asociada a CRISPR, familia Csn1 [Roseburia intestinalis L1-82]
gi|323142435|ref|ZP_08077256.1|, proteína asociada a CRISPR, familia Csn1 [Phascolarctobacterium succinatutens YIT 12067]
gi|310780384|ref|YP_003968716.1|, proteína asociada a CRISPR, familia Csn1 [Ilyobacter polytropus DSM 2926]
gi|336393381|ref|ZP_08574780.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus coryniformis subsp. torquens KCTC 3535...
gi|189485225|ref|YP_001956166.1|, proteína Csn1 asociada a CRISPR [no cultivado Termite grupo 1 bacteria filotipo Rs-D17]
gi|220930482|ref|YP_002507391.1|, proteína asociada a CRISPR, familia Csn1 [Clostridium cellulolyticum H10]
gi|157415744|ref|YP_001483000.1|, proteína hipotética C8J_1425 [Campylobacter jejuni subsp. jejuni 81116]
gi|218563121|ref|YP_002344900.1|, proteína asociada a CRISPR [Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819]
gi|153952471|ref|YP_001398821.1|, proteína de la familia Cas5e asociada a CRISPR [Campylobacter jejuni subsp. doylei 269.97]
gi|313144862|ref|ZP_07807055.1|, proteína a CRISPR [Helicobacter cinaedi CCUG 18818]
gi|253828136|ref|ZP_04871021.1|, proteína asociada a CRISPR, familia Csn1 [Helicobacter canadensis MIT 98-5491]
gi|291276265|ref|YP_003516037.1|, proteína asociada asociada a CRISPR [Helicobacter mustelae 12198]
gi|345577901|ref|NP_907605.1|, proteína hipotética WS1445 [Wolinella succinogenes DSM 1740]

FIG. 8A

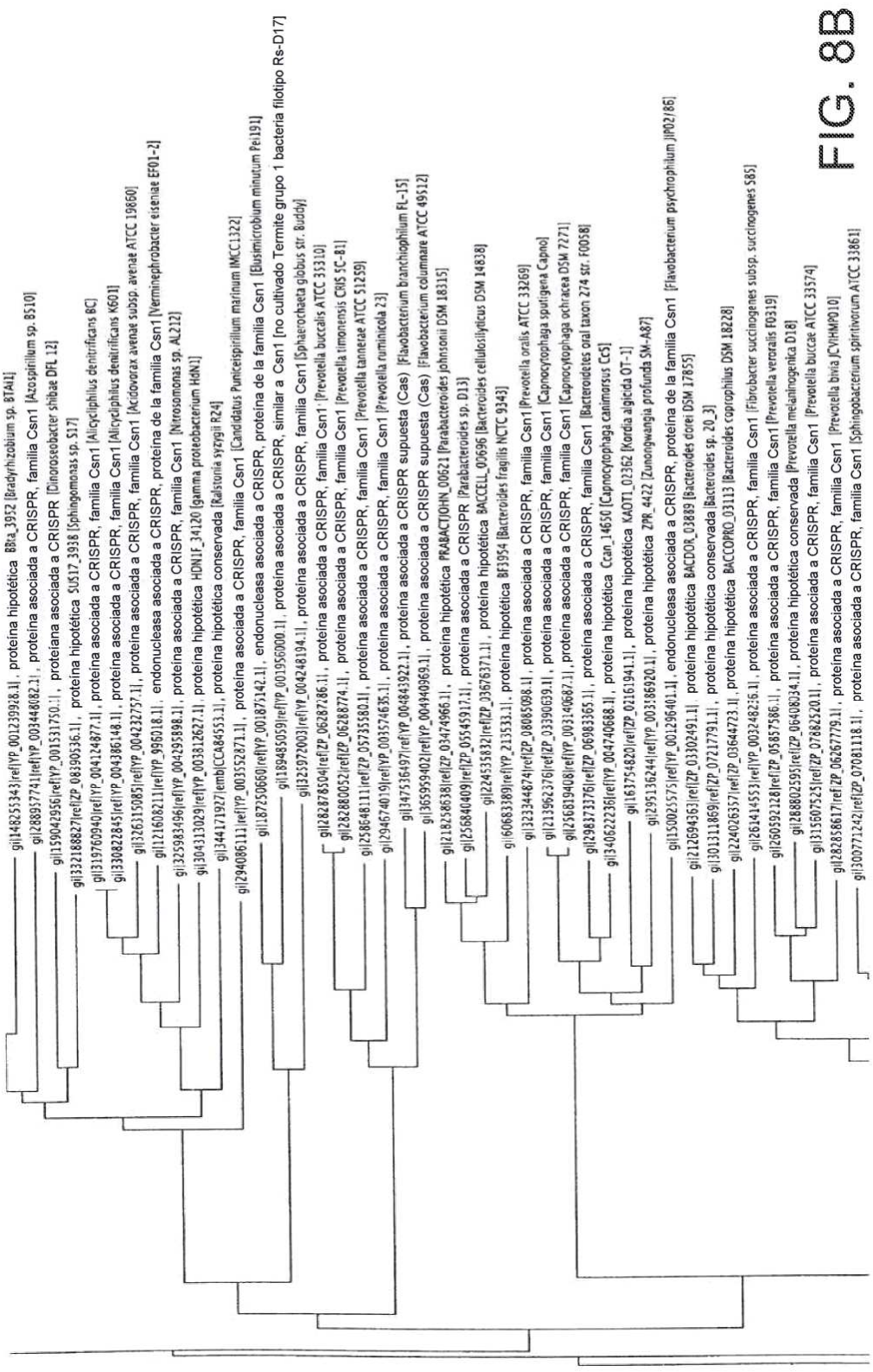

gi|148255343|ref|YP_001239938.1|, proteína hipotética BBta_3952 [Bradyrhizobium sp. BTAi1]
gi|288957741|ref|YP_003448082.1|, proteína asociada a CRISPR, familia Csn1 [Azospirillum sp. B510]
gi|159042956|ref|YP_001531750.1|, proteiana asociada a CRISPR [Dinoroseobacter shibae DFL 12]
gi|332188827|ref|ZP_08390536.1|, proteína hipotética SU17_3938 [Sphingomonas sp. S17]
gi|319760940|ref|YP_004124877.1|, proteína asociada a CRISPR, familia Csn1 [Alicycliphilus denitrificans BC]
gi|330822845|ref|YP_004386148.1|, proteína asociada a CRISPR, familia Csn1 [Alicycliphilus denitrificans K601]
gi|326315085|ref|YP_004232757.1|, proteína asociada a CRISPR, familia Csn1 [Acidovorax avenae subsp. avenae ATCC 19860]
gi|121608211|ref|YP_995018.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Verminephrobacter eiseniae EF01-2]
gi|325983496|ref|YP_004295898.1|, proteína asociada a CRISPR, familia Csn1 [Nitrosomonas sp. AL212]
gi|304313029|ref|YP_003812627.1|, proteína hipotética HDN1F_34120 [gamma proteobacterium HdN1]
gi|344171927|emb|CCA84553.1|, proteína hipotética conservada [Ralstonia syzygii R24]
gi|294086111|ref|YP_003552871.1|, proteína asociada a CRISPR, familia Csn1 [Candidatus Puniceispirillum marinum IMCC1322]
gi|187250650|ref|YP_001875142.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Busmicrobium minutum Pei191]
gi|189485059|ref|YP_001956000.1|, proteína asociada a CRISPR, similar a Csn1 [no cultivado Termite grupo 1 bacteria filotipo Rs-D17]
gi|325972003|ref|YP_004248194.1|, proteína asociada a CRISPR, familia Csn1 [Sphaerochaeta globus str. Buddy]
gi|282878504|ref|ZP_06287286.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella buccalis ATCC 35310]
gi|282880052|ref|ZP_06288774.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella timonensis CRIS 5C-B1]
gi|258648111|ref|ZP_05735580.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella tannerae ATCC 51259]
gi|294674019|ref|YP_003574635.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella ruminicola 23]
gi|347316497|ref|YP_004843922.1|, proteína asociada a CRISPR supuesta (Cas) [Flavobacterium branchiophilum FL-15]
gi|365959402|ref|YP_004940969.1|, proteína asociada a CRISPR supuesta (Cas) [Flavobacterium columnare ATCC 49512]
gi|218258638|ref|ZP_03474966.1|, proteína hipotética PRABACTJOHN_00621 [Parabacteroides johnsonii DSM 18315]
gi|256840409|ref|ZP_05545917.1|, proteína asociada a CRISPR [Parabacteroides sp. D13]
gi|224535832|ref|ZP_03676371.1|, proteína hipotética BACCELL_00696 [Bacteroides cellulosilyticus DSM 14838]
gi|60683389|ref|YP_213533.1|, proteína hipotética BF3954 [Bacteroides fragilis NCTC 9343]
gi|323344874|ref|ZP_08085098.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella oralis ATCC 33269]
gi|213962376|ref|ZP_03390039.1|, proteína asociada a CRISPR, familia Csn1 [Capnocytophaga sputigena Capno]
gi|256819408|ref|YP_003140687.1|, proteína asociada a CRISPR, familia Csn1 [Capnocytophaga ochracea DSM 7271]
gi|298373376|ref|ZP_06983365.1|, proteína asociada a CRISPR, familia Csn1 [Bacteroidetes oral taxon 274 str. F0058]
gi|340622236|ref|YP_004740688.1|, proteína hipotética Ccan_14650 [Capnocytophaga canimorsus Cc5]
gi|161754820|ref|ZP_02161941.1|, proteína hipotética KAOT1_02362 [Kordia algicida OT-1]
gi|295136244|ref|YP_003586920.1|, proteína hipotética ZPR_4422 [Zunongwangia profunda SM-A87]
gi|150025575|ref|YP_001296401.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Flavobacterium psychrophilum JIP02/86]
gi|212694363|ref|ZP_03302491.1|, proteína hipotética BACDOR_03689 [Bacteroides dorei DSM 17855]
gi|301311869|ref|ZP_07217791.1|, proteína hipotética conservada [Bacteroides sp. 20_3]
gi|224026357|ref|ZP_03644723.1|, proteína hipotética BACCOPRO_03113 [Bacteroides coprophilus DSM 18228]
gi|261414553|ref|YP_003248236.1|, proteína asociada a CRISPR, familia Csn1 [Fibrobacter succinogenes subsp. succinogenes S85]
gi|260592128|ref|ZP_05857586.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella veroralis F0319]
gi|288800259|ref|ZP_06408434.1|, proteína hipotética conservada [Prevotella melaninogenica D18]
gi|315607523|ref|ZP_07882520.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella buccae ATCC 33574]
gi|282856617|ref|ZP_06267779.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella bivia JCVIHMP010]
gi|300771242|ref|ZP_07081118.1|, proteína asociada a CRISPR, familia Csn1 [Sphingobacterium spiritivorum ATCC 33861]

FIG. 8B


gi|148255343|ref|YP_001239928.1|, proteína hipotética BBta_3952 [Bradyrhizobium sp. BTAi1]
gi|288957741|ref|YP_003448082.1|, proteína asociada a CRISPR, familia Csn1 [Azospirillum sp. B510]
gi|159042956|ref|YP_001531750.1|, proteína asociada a CRISPR [Dinoroseobacter shibae DFL 12]
gi|332186827|ref|ZP_08390536.1|, proteína hipotética SJ517_3938 [Sphingomonas sp. S17]
gi|319760940|ref|YP_004124877.1|, proteína asociada a CRISPR, familia Csn1 [Alicycliphilus denitrificans BC]
gi|330822845|ref|YP_004386148.1|, proteína asociada a CRISPR, familia Csn1 [Alicycliphilus denitrificans K601]
gi|326511685|ref|YP_004232757.1|, proteína asociada a CRISPR, familia Csn1 [Acidovorax avenae subsp. avenae ATCC 19860]
gi|121608211|ref|YP_996018.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Verminephrobacter eiseniae EF01-2]
gi|325983496|ref|YP_004295898.1|, proteína asociada a CRISPR, familia Csn1 [Nitrosomonas sp. AL212]
gi|304313029|ref|YP_003812627.1|, proteína hipotética HDN1F_34120 [gamma proteobacterium HdN1]
gi|344171927|emb|CCA84553.1|, proteína hipotética conservada [Ralstonia syzygii R24]
gi|294086111|ref|YP_003552871.1|, proteína asociada a CRISPR, familia Csn1 [Candidatus Puniceispirillum marinum IMCC1322]
gi|187250960|ref|YP_001875142.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Elusimicrobium minutum Pei191]
gi|189485059|ref|YP_001956000.1|, proteína asociada a CRISPR, similar a Csn1 [no cultivado Termite grupo 1 bacteria filotipo Rs-D17]
gi|325972003|ref|YP_004248194.1|, proteína asociada a CRISPR, familia Csn1 [Spherochaeta globus str. Buddy]
gi|282877504|ref|ZP_06287266.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella buccalis ATCC 35310]
gi|282880052|ref|ZP_06288774.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella timonensis CRIS 5C-B1]
gi|258648111|ref|ZP_05735580.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella tannerae ATCC 51259]
gi|294674019|ref|YP_003574635.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella ruminicola 23]
gi|347516457|ref|YP_004843922.1|, proteína asociada a CRISPR supuesta (Cas) [Flavobacterium branchiophilum FL-15]
gi|365959402|ref|YP_004940969.1|, proteína asociada a CRISPR supuesta (Cas) [Flavobacterium columnare ATCC 49512]
gi|218258638|ref|ZP_03474966.1|, proteína hipotética PRABACTJOHN_00621 [Parabacteroides johnsonii DSM 18315]
gi|256840409|ref|ZP_05545917.1|, proteína asociada a CRISPR [Parabacteroides sp. D13]
gi|224535832|ref|ZP_03676171.1|, proteína hipotética BACCELL_00596 [Bacteroides cellulosilyticus DSM 14838]
gi|60683389|ref|YP_213531.1|, proteína hipotética BF3954 [Bacteroides fragilis NCTC 9343]
gi|323344874|ref|ZP_08085098.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella oralis ATCC 33269]
gi|213962376|ref|ZP_03390639.1|, proteína asociada a CRISPR, familia Csn1 [Capnocytophaga sputigena Capno]
gi|256819408|ref|YP_003140687.1|, proteína asociada a CRISPR, familia Csn1 [Capnocytophaga ochracea DSM 7271]
gi|298373376|ref|ZP_06983365.1|, proteína asociada a CRISPR, familia Csn1 [Bacteroidetes oral taxon 274 str. F0058]
gi|340622236|ref|YP_004740688.1|, proteína hipotética Ccan_14650 [Capnocytophaga canimorsus Cc5]
gi|315375420|ref|ZP_02161941.1|, proteína hipotética KAOT1_02362 [Kordia algicida OT-1]
gi|295113244|ref|YP_003568920.1|, proteína hipotética ZPR_4422 [Zunongwangia profunda SM-A87]
gi|150025375|ref|YP_001296401.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Flavobacterium psychrophilum JIP02/86]
gi|212694363|ref|ZP_03302491.1|, proteína hipotética BACCOR_03689 [Bacteroides dorei DSM 17855]
gi|301311869|ref|ZP_07217791.1|, proteína hipotética conservada [Bacteroides sp. 20_3]
gi|224026357|ref|ZP_03644723.1|, proteína hipotética BACCOPRO_03113 [Bacteroides coprophilus DSM 18228]
gi|261414533|ref|YP_003248236.1|, proteína asociada a CRISPR, familia Csn1 [Fibrobacter succinogenes subsp. succinogenes S85]
gi|260592128|ref|ZP_05857386.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella veroralis F0319]
gi|288802595|ref|ZP_06408034.1|, proteína hipotética conservada [Prevotella melaninogenica D18]
gi|315507525|ref|ZP_07882520.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella buccae ATCC 33574]
gi|282859617|ref|ZP_06267779.1|, proteína asociada a CRISPR, familia Csn1 [Prevotella bivia JCVIHMP010]

FIG. 8C

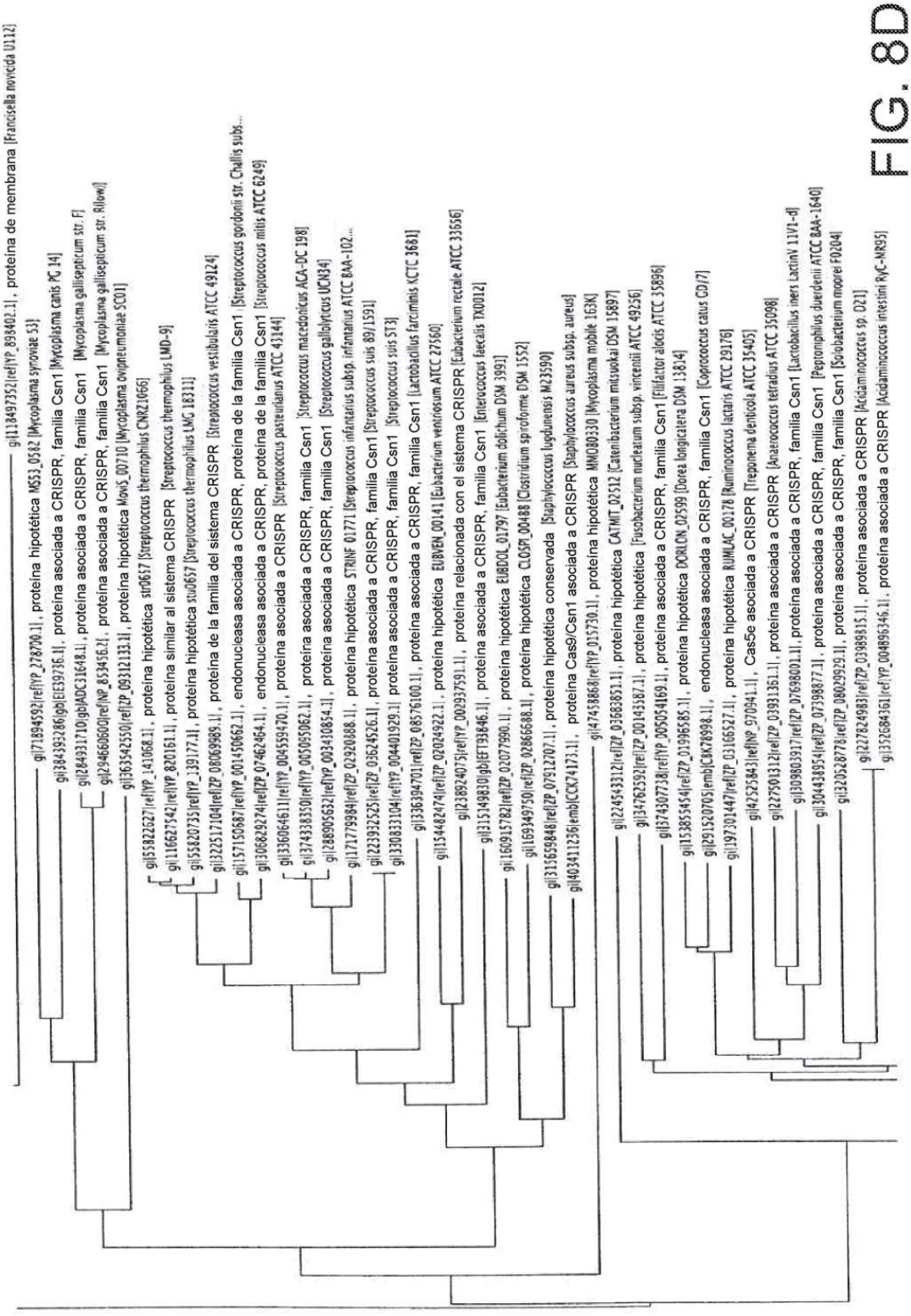

gi|113497352|ref|YP_893402.1|, proteína de membrana [Francisella novicida U112]
gi|71894592|ref|YP_278700.1|, proteína hipotética MS53_0582 [Mycoplasma synoviae 53]
gi|384393286|gb|EIE39736.1|, proteína asociada a CRISPR, familia Csn1 [Mycoplasma canis PG 14]
gi|284931710|gb|ADC31648.1|, proteína asociada a CRISPR, familia Csn1 [Mycoplasma gallisepticum str. F]
gi|294660600|ref|NP_853456.2|, proteína asociada a CRISPR, familia Csn1 [Mycoplasma gallisepticum str. Rlow]
gi|363542550|ref|ZP_09312133.1|, proteína hipotética MovS_0710 [Mycoplasma ovipneumoniae SC01]
gi|558226271|ref|YP_001461068.1|, proteína hipotética str0617 [Streptococcus thermophilus CNRZ1066]
gi|116627542|ref|YP_820161.1|, proteína similar al sistema CRISPR [Streptococcus thermophilus LMD-9]
gi|558220735|ref|YP_139177.1|, proteína hipotética stu0617 [Streptococcus thermophilus LMG 18311]
gi|322517104|ref|ZP_08069989.1|, proteína de la familia del sistema CRISPR [Streptococcus vestibularis ATCC 49124]
gi|157150687|ref|YP_001450662.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1. [Streptococcus gordonii str. Challis subs...
gi|306829274|ref|ZP_07462464.1|, endonucleasa asociada a CRISPR, proteína de la familia Csn1 [Streptococcus mitis ATCC 6249]
gi|336064611|ref|YP_004559470.1|, proteína asociada a CRISPR [Streptococcus pasteurianus ATCC 43144]
gi|374338350|ref|YP_005055062.1|, proteína asociada a CRISPR, familia Csn1 [Streptococcus macedonicus ACA-DC 198]
gi|288905632|ref|YP_003430654.1|, proteína asociada a CRISPR, familia Csn1 [Streptococcus gallolyticus UCN34]
gi|171779984|ref|ZP_02920888.1|, proteína hipotética STRINF_01771 [Streptococcus infantarius subsp. infantarius ATCC BAA-102...
gi|223932525|ref|ZP_03624526.1|, proteína asociada a CRISPR, familia Csn1 [Streptococcus suis 89/1591]
gi|330833104|ref|YP_004401929.1|, proteína asociada a CRISPR, familia Csn1 [Streptococcus suis ST3]
gi|336394701|ref|ZP_08576100.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus farciminis KCTC 3681]
gi|154482474|ref|ZP_02024922.1|, proteína hipotética EUBVEN_00141 [Eubacterium ventriosum ATCC 27560]
gi|238924075|ref|YP_002937591.1|, proteína relacionada con el sistema CRISPR [Eubacterium rectale ATCC 33656]
gi|315149830|gb|EFT93846.1|, proteína asociada a CRISPR, familia Csn1 [Enterococcus faecalis TX0012]
gi|160915782|ref|ZP_02077990.1|, proteína hipotética EUBDOL_01797 [Eubacterium dolichum DSM 3991]
gi|169349750|ref|ZP_02866688.1|, proteína hipotética CLOSPI_00488 [Clostridium spiroforme DSM 1552]
gi|315659848|ref|ZP_07912707.1|, proteína hipotética conservada [Staphylococcus lugdunensis M23590]
gi|40341236|emb|CCK74173.1|, proteína Cas9/Csn1 asociada a CRISPR [Staphylococcus aureus subsp. aureus]
gi|47458868|ref|YP_015730.1|, proteína hipotética MMOB0330 [Mycoplasma mobile 163K]
gi|224543312|ref|ZP_03683851.1|, proteína hipotética CATMIT_02512 [Catenibacterium mitsuokai DSM 15897]
gi|347623932|ref|ZP_00143587.1|, proteína hipotética [Fusobacterium nucleatum subsp. vincentii ATCC 49256]
gi|374307738|ref|YP_005054869.1|, proteína asociada a CRISPR, familia Csn1 [Filifactor alocis ATCC 35896]
gi|153855454|ref|ZP_01996585.1|, proteína hipotética DORLON_02599 [Dorea longicatena DSM 13814]
gi|291520705|emb|CBK78998.1|, endonucleasa asociada a CRISPR, familia Csn1 [Coprococcus catus GD/7]
gi|197301447|ref|ZP_03166527.1|, proteína hipotética RUMLAC_00178 [Ruminococcus lactaris ATCC 29176]
gi|342325843|ref|NP_970941.1|, Cas5e asociada a CRISPR [Treponema denticola ATCC 35405]
gi|227501312|ref|ZP_03931361.1|, proteína asociada a CRISPR [Anaerococcus tetradius ATCC 35098]
gi|309803917|ref|ZP_07698001.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus iners LactinV 11V1-d]
gi|304438954|ref|ZP_07398877.1|, proteína asociada a CRISPR, familia Csn1 [Peptoniphilus duerdenii ATCC BAA-1640]
gi|320528778|ref|ZP_08029929.1|, proteína asociada a CRISPR, familia Csn1 [Solobacterium moorei F0204]
gi|227824983|ref|ZP_03989815.1|, proteína asociada a CRISPR [Acidaminococcus sp. D21]
gi|356684361|ref|YP_004896346.1|, proteína asociada a CRISPR [Acidaminococcus intestini RyC-MR95]

FIG. 8D

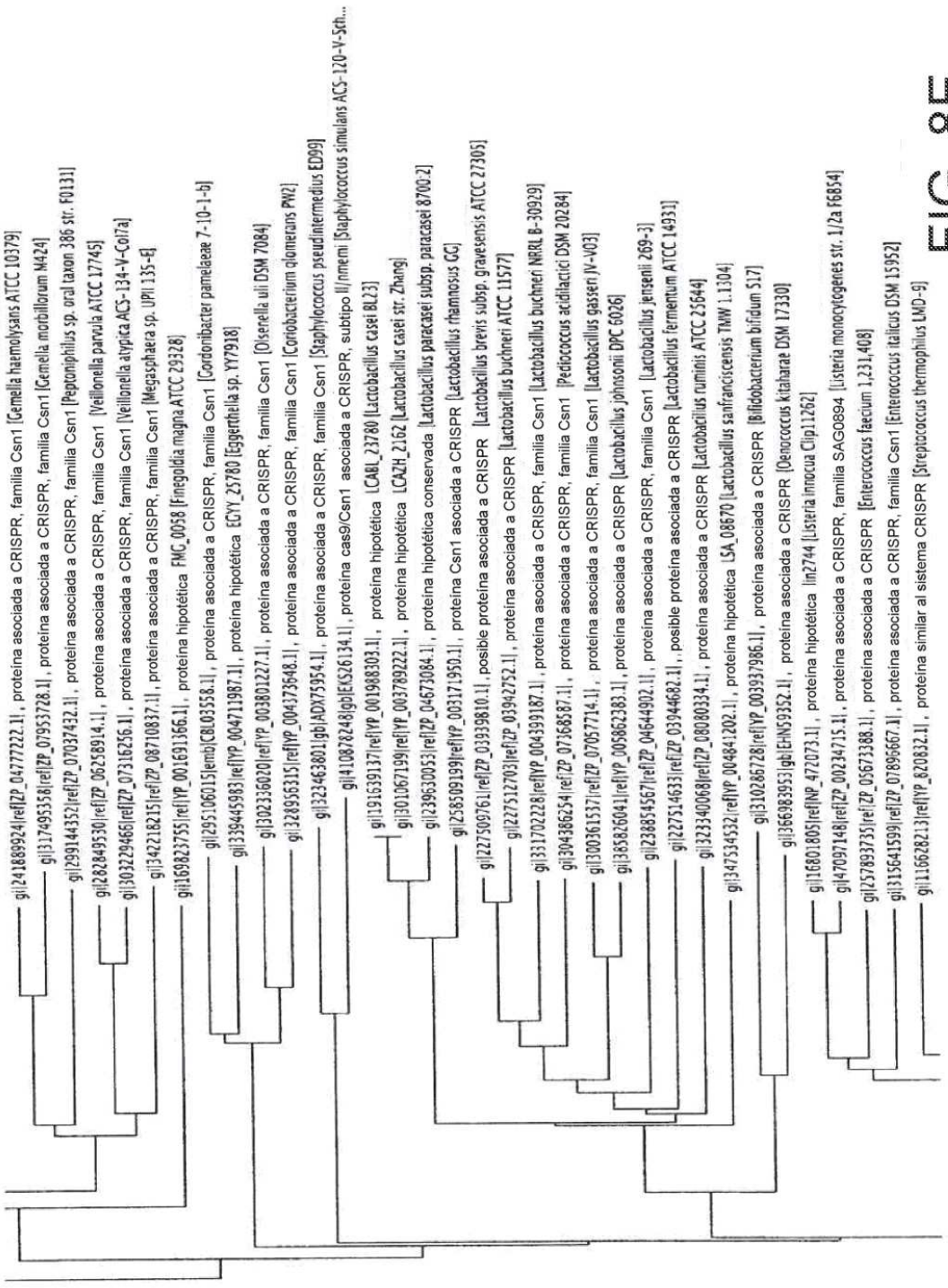

gi|241889924|ref|ZP_04777222.1|, proteína asociada a CRISPR, familia Csn1 [Gemella haemolysans ATCC 10379]
gi|317495358|ref|ZP_07953728.1|, proteína asociada a CRISPR, familia Csn1 [Gemella morbillorum M424]
gi|299144352|ref|ZP_07037432.1|, proteína asociada a CRISPR, familia Csn1 [Peptoniphilus sp. oral taxon 386 str. F0131]
gi|282849530|ref|ZP_06258914.1|, proteína asociada a CRISPR, familia Csn1 [Veillonella parvula ATCC 17745]
gi|303229466|ref|ZP_07316256.1|, proteína asociada a CRISPR, familia Csn1 [Veillonella atypica ACS-134-V-Col7a]
gi|342218215|ref|ZP_08710837.1|, proteína asociada a CRISPR, familia Csn1 [Megasphaera sp. UPII 135-E]
gi|169823755|ref|YP_001691366.1|, proteína hipotética FMG_0058 [Finegoldia magna ATCC 29328]
gi|295106015|emb|CBL03558.1|, proteína asociada a CRISPR, familia Csn1 [Corynebacter pamelaeae 7-10-1-b]
gi|339445983|ref|YP_004711987.1|, proteína hipotética EGYY_25780 [Eggerthella sp. YY7918]
gi|302336020|ref|YP_003801227.1|, proteína asociada a CRISPR, familia Csn1 [Olsenella uli DSM 7084]
gi|328956315|ref|YP_004373648.1|, proteína asociada a CRISPR, familia Csn1 [Coriobacterium glomerans PW2]
gi|323463801|gb|ADX75954.1|, proteína asociada a CRISPR, familia Csn1 [Staphylococcus pseudintermedius ED99]
gi|410878248|gb|EKS26134.1|, proteína cas9/Csn1 asociada a CRISPR, subtipo II/nmemi [Staphylococcus simulans ACS-120-V-Sch...
gi|319639137|ref|YP_001988303.1|, proteína hipotética LCABL_23780 [Lactobacillus casei BL23]
gi|301067199|ref|YP_003789222.1|, proteína hipotética LCAZH_2162 [Lactobacillus casei str. Zhang]
gi|239630053|ref|ZP_04673084.1|, proteína hipotética conservada [Lactobacillus paracasei subsp. paracasei 8700:2]
gi|258509199|ref|YP_003171950.1|, proteína Csn1 asociada a CRISPR [Lactobacillus rhamnosus GG]
gi|227509761|ref|ZP_03938810.1|, posible proteína asociada a CRISPR [Lactobacillus brevis subsp. gravesensis ATCC 27305]
gi|227512703|ref|ZP_03942752.1|, proteína asociada a CRISPR [Lactobacillus buchneri ATCC 11577]
gi|331702228|ref|YP_004399187.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus buchneri NRRL B-30929]
gi|304386254|ref|ZP_07368587.1|, proteína asociada a CRISPR, familia Csn1 [Pediococcus acidilactici DSM 20284]
gi|300361537|ref|ZP_07057714.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus gasseri JV-V03]
gi|365826041|ref|YP_005062383.1|, proteína asociada a CRISPR [Lactobacillus johnsonii DPC 6026]
gi|238854567|ref|ZP_04644902.1|, proteína asociada a CRISPR, familia Csn1 [Lactobacillus jensenii 269-3]
gi|227514633|ref|ZP_03944682.1|, posible proteína asociada a CRISPR [Lactobacillus fermentum ATCC 14933]
gi|323340068|ref|ZP_08080334.1|, proteína asociada a CRISPR [Lactobacillus ruminis ATCC 25644]
gi|347534532|ref|YP_004841202.1|, proteína hipotética LSA_08670 [Lactobacillus sanfranciscensis TMW 1.1304]
gi|310286728|ref|YP_003937986.1|, proteína asociada a CRISPR [Bifidobacterium bifidum S17]
gi|366983953|gb|EHN59352.1|, proteína asociada a CRISPR [Oenococcus kitaharae DSM 17330]
gi|116801805|ref|NP_472073.1|, proteína hipotética lin2744 [Listeria innocua Clip11262]
gi|47097148|ref|ZP_00234715.1|, proteína asociada a CRISPR, familia SAG0894 [Listeria monocytogenes str. 1/2a F6854]
gi|257893735|ref|ZP_05673388.1|, proteína asociada a CRISPR [Enterococcus faecium 1,231,408]
gi|315641599|ref|ZP_07896667.1|, proteína asociada a CRISPR, familia Csn1 [Enterococcus italicus DSM 15952]
gi|116628213|ref|YP_820832.1|, proteína similar al sistema CRISPR [Streptococcus thermophilus LMD-9]

FIG. 8E

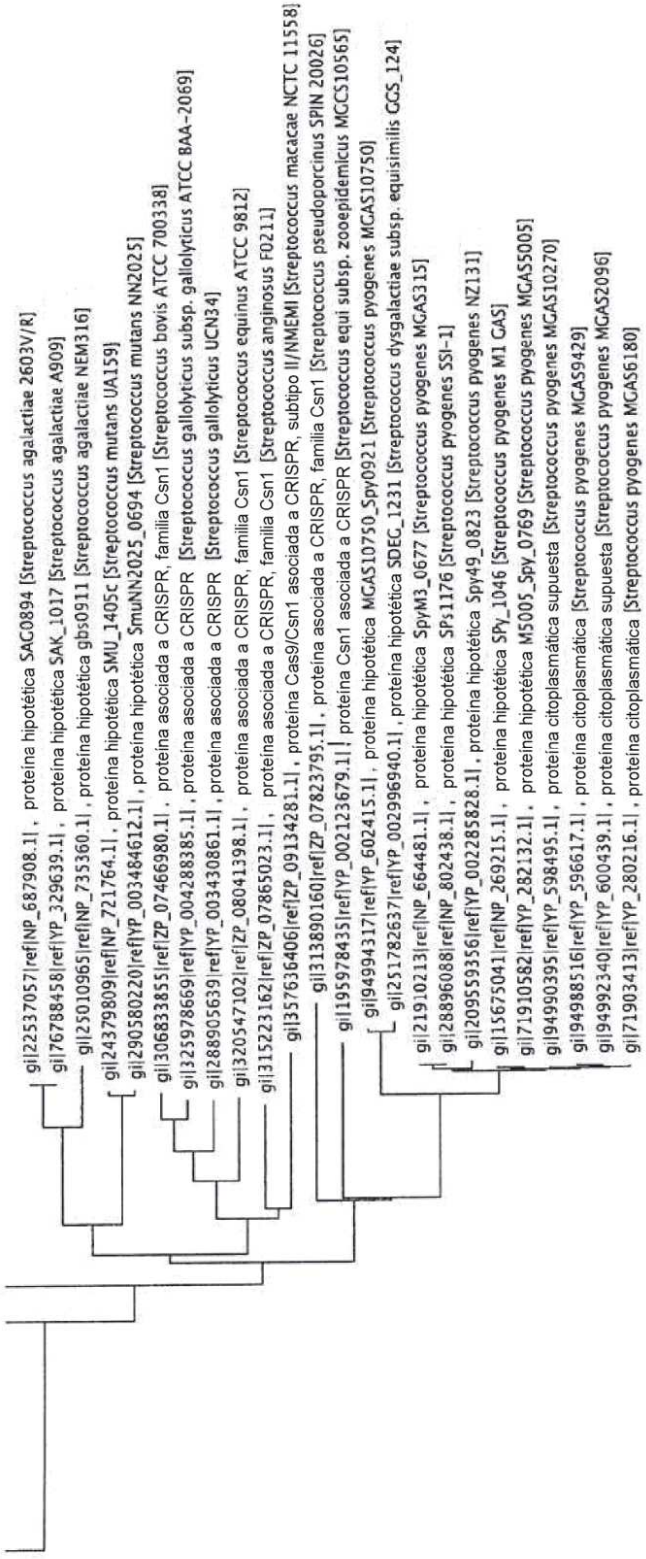

gi|22537057|ref|NP_687908.1| , proteína hipotética SAG0894 [Streptococcus agalactiae 2603V/R]
gi|76788458|ref|YP_329639.1| , proteína hipotética SAK_1017 [Streptococcus agalactiae A909]
gi|25010965|ref|NP_735360.1| , proteína hipotética gbs0911 [Streptococcus agalactiae NEM316]
gi|24379809|ref|NP_721764.1| , proteína hipotética SMU_1405c [Streptococcus mutans UA159]
gi|290580220|ref|YP_003484612.1| , proteína hipotética SmuNN2025_0694 [Streptococcus mutans NN2025]
gi|306833855|ref|ZP_07466980.1| , proteína asociada a CRISPR, familia Csn1 [Streptococcus bovis ATCC 700338]
gi|325978669|ref|YP_004288385.1| , proteína asociada a CRISPR [Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069]
gi|288905639|ref|YP_003430861.1| , proteína asociada a CRISPR [Streptococcus gallolyticus UCN34]
gi|320547102|ref|ZP_08041398.1| , proteína asociada a CRISPR, familia Csn1 [Streptococcus equinus ATCC 9812]
gi|315223162|ref|ZP_07865023.1| , proteína asociada a CRISPR, familia Csn1 [Streptococcus anginosus F0211]
gi|357636406|ref|ZP_09134281.1| , proteína Cas9/Csn1 asociada a CRISPR, subtipo II/NMEMI [Streptococcus macacae NCTC 11558]
gi|313890160|ref|ZP_07823795.1| , proteína asociada a CRISPR, familia Csn1 [Streptococcus pseudoporcinus SPIN 20026]
gi|195978435|ref|YP_002123679.1| , proteína Csn1 asociada a CRISPR [Streptococcus equi subsp. zooepidemicus MGCS10565]
gi|94994317|ref|YP_602415.1| , proteína hipotética MGAS10750_Spy0921 [Streptococcus pyogenes MGAS10750]
gi|251782637|ref|YP_002996940.1| , proteína hipotética SDEG_1231 [Streptococcus dysgalactiae subsp. equisimilis GGS_124]
gi|21910213|ref|NP_664481.1| , proteína hipotética SpyM3_0677 [Streptococcus pyogenes MGAS315]
gi|28896088|ref|NP_802438.1| , proteína hipotética SPs1176 [Streptococcus pyogenes SSI-1]
gi|209559356|ref|YP_002285828.1| , proteína hipotética Spy49_0823 [Streptococcus pyogenes NZ131]
gi|15675041|ref|NP_269215.1| , proteína hipotética SPy_1046 [Streptococcus pyogenes M1 GAS]
gi|71910582|ref|YP_282132.1| , proteína hipotética M5005_Spy_0769 [Streptococcus pyogenes MGAS5005]
gi|94990395|ref|YP_598495.1| , proteína citoplasmática supuesta [Streptococcus pyogenes MGAS10270]
gi|94988516|ref|YP_596617.1| , proteína citoplasmática [Streptococcus pyogenes MGAS9429]
gi|94992340|ref|YP_600439.1| , proteína citoplasmática supuesta [Streptococcus pyogenes MGAS2096]
gi|71903413|ref|YP_280216.1| , proteína citoplasmática [Streptococcus pyogenes MGAS6180]

FIG. 8F

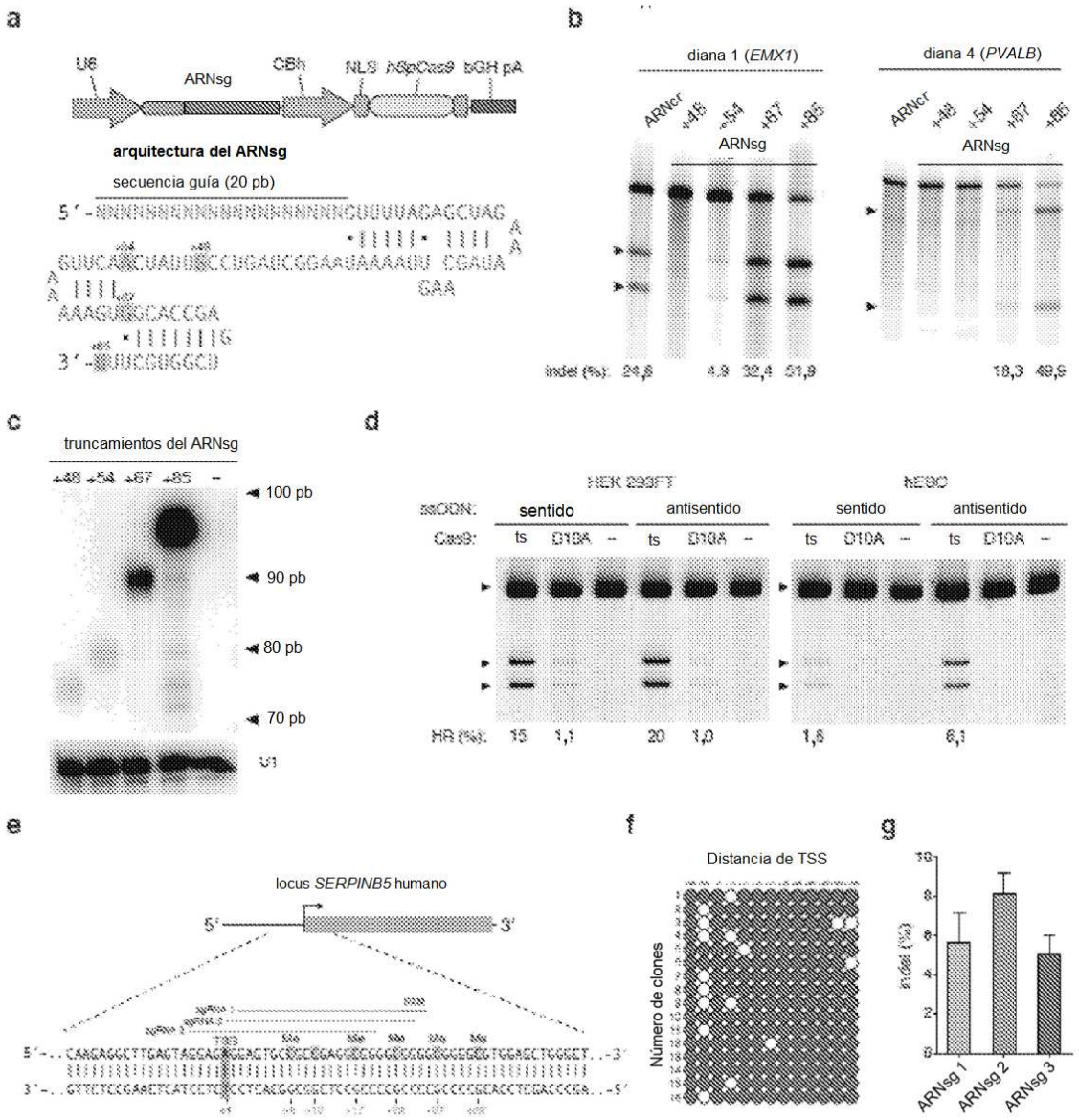

a
U6
ARNsg
CBh
NLS
bGH pA
arquitectura del ARNsg
secuencia guía (20 pb)
b
diana 1 (EMX1)
diana 4 (PVALB)
ARNcr
+48
+54
+67
+85
ARNsg
Indel (%): 24,8 4,9 32,4 51,9
18,3 48,9
c
truncamientos del ARNsg
+48 +54 +67 +85 –
100 pb
90 pb
80 pb
70 pb
U1
d
HEK 293FT
NESC
ssODN:
sentido
antisentido
Cas9:
ts
D10A
–
HR (%): 15 1,1 20 1,0 1,8 6,1
e
locus SERPINB5 humano
5'
3'
f
Distancia de TSS
Número de clones
g
Indel (%)
ARNsg 1
ARNsg 2
ARNsg 3

FIG. 9

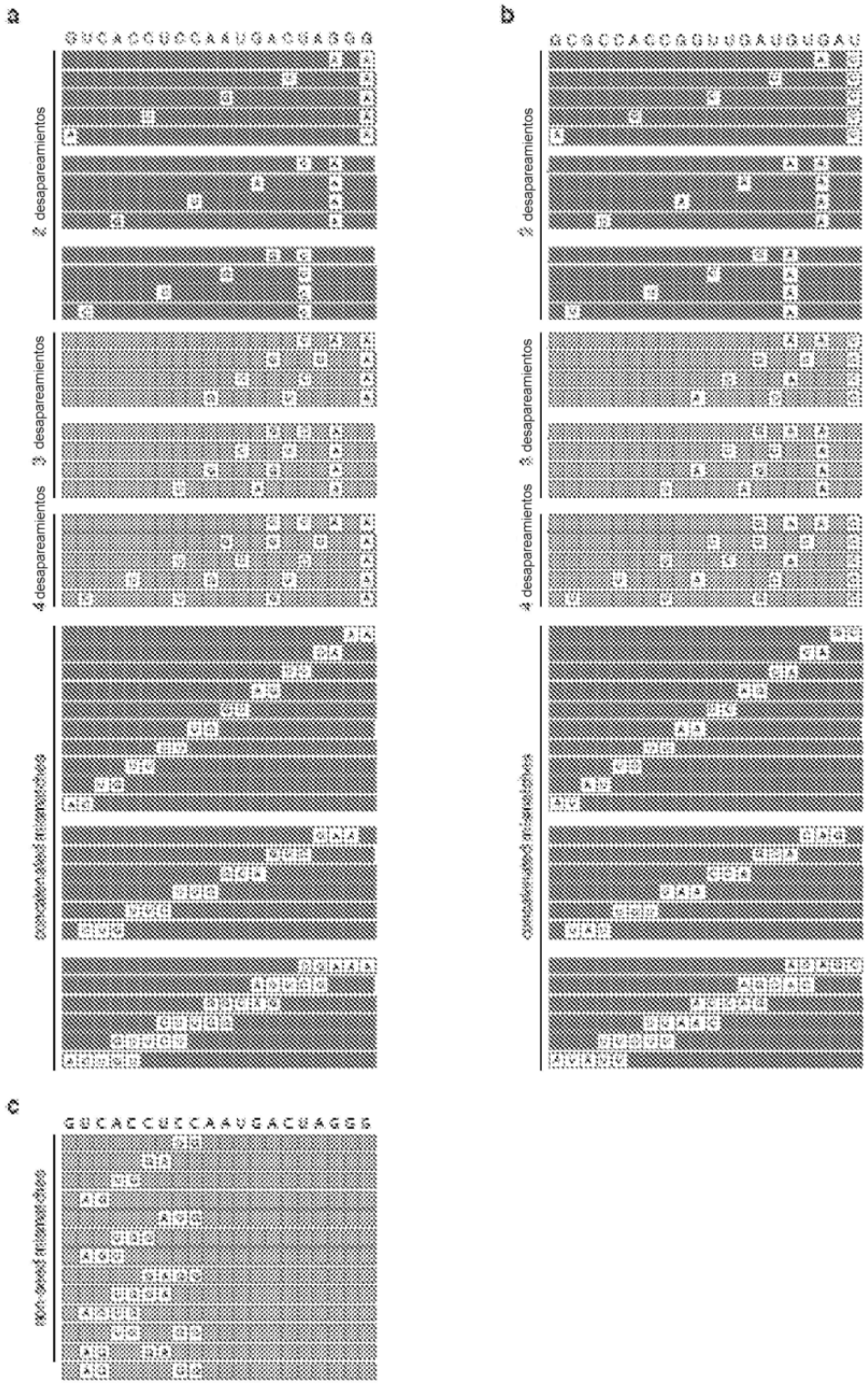

a
GUCACCUCCAAUGACGASGG
2 desapareamientos
3 desapareamientos
4 desapareamientos
b
GCGCCACCGGUUGAUGUGAU
2 desapareamientos
3 desapareamientos
4 desapareamientos
c
GUCACCUCCAAUGACUAGGG

FIG. 10

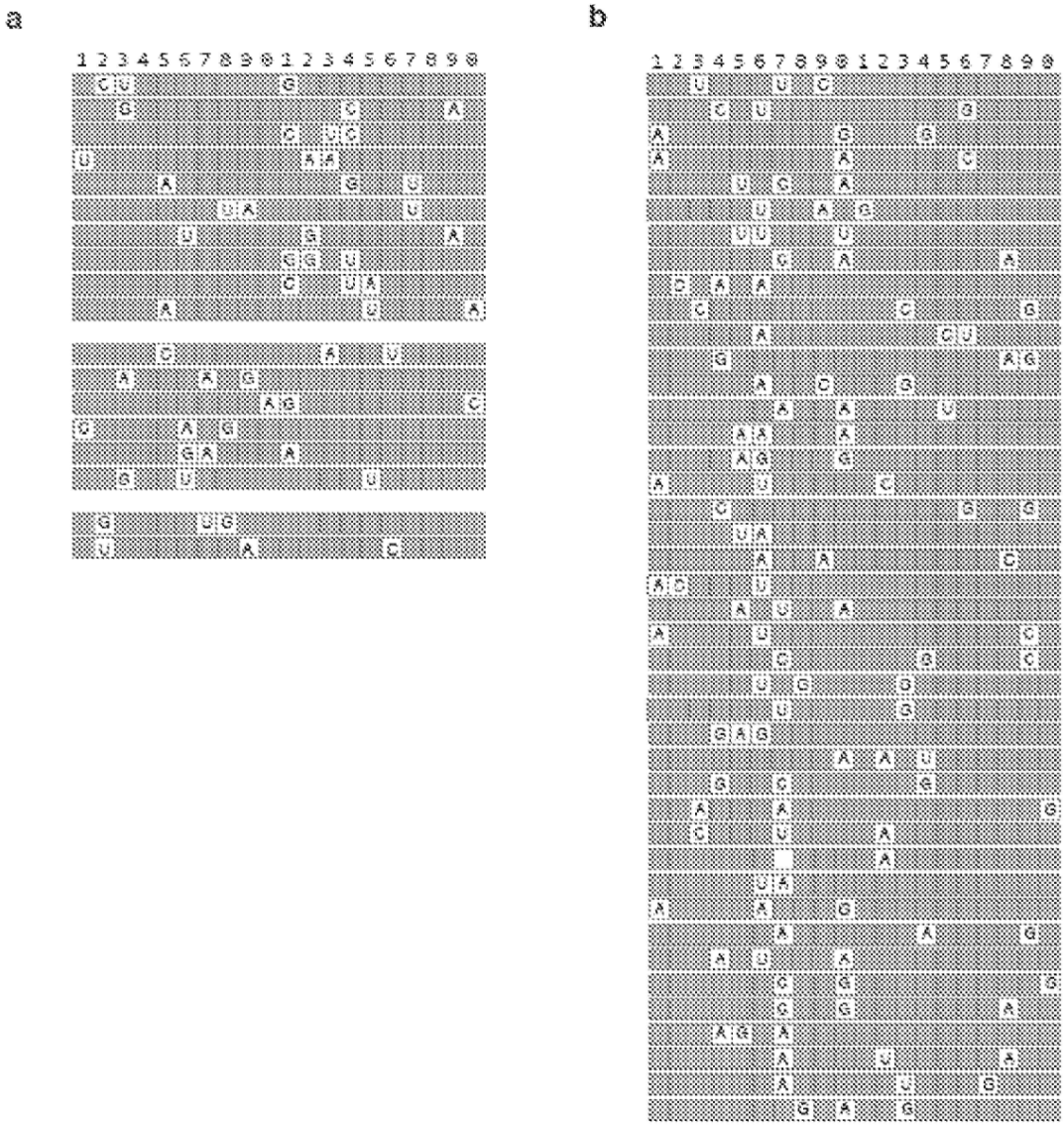


FIG. 11

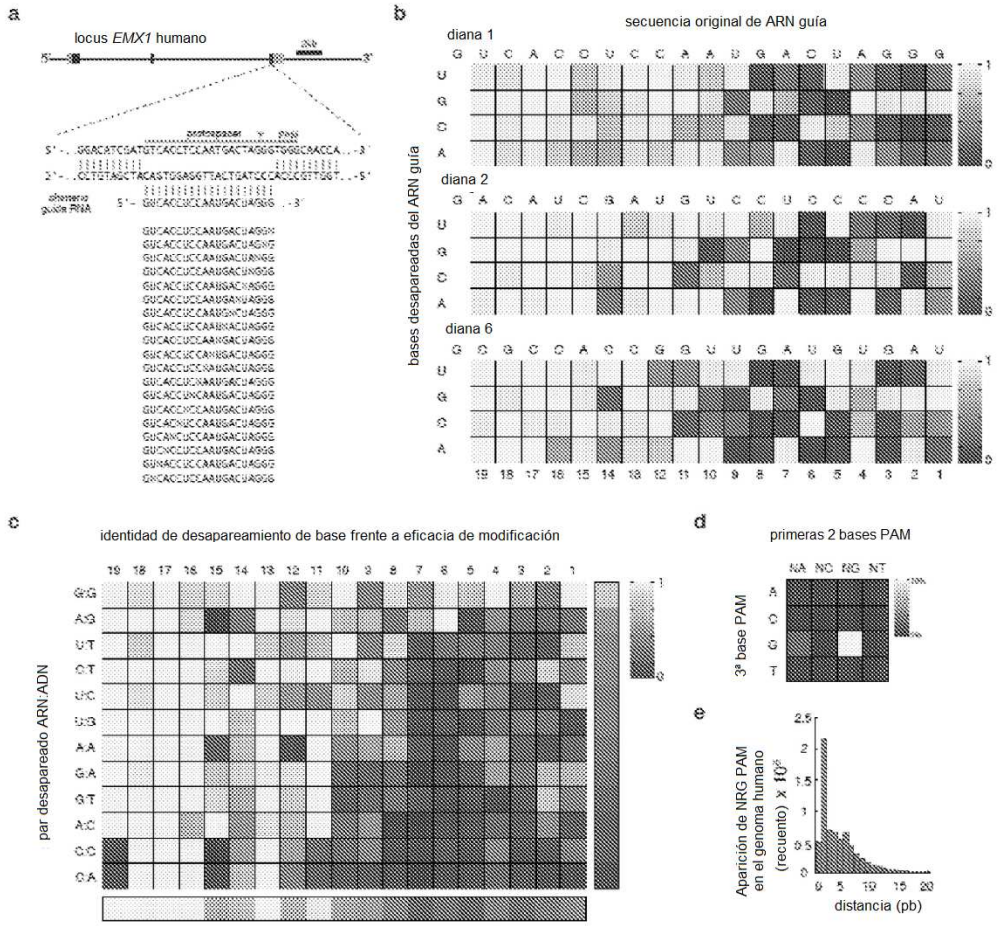

a
locus *EMX1* humano
b
secuencia original de ARN guía
diana 1
diana 2
diana 6
bases desapareadas del ARN guía
c
identidad de desapareamiento de base frente a eficacia de modificación
par desapareado ARN:ADN
d
primeras 2 bases PAM
3ª base PAM
e
Aparición de NRG PAM en el genoma humano (recuento) x 10^6
distancia (pb)

FIG. 12

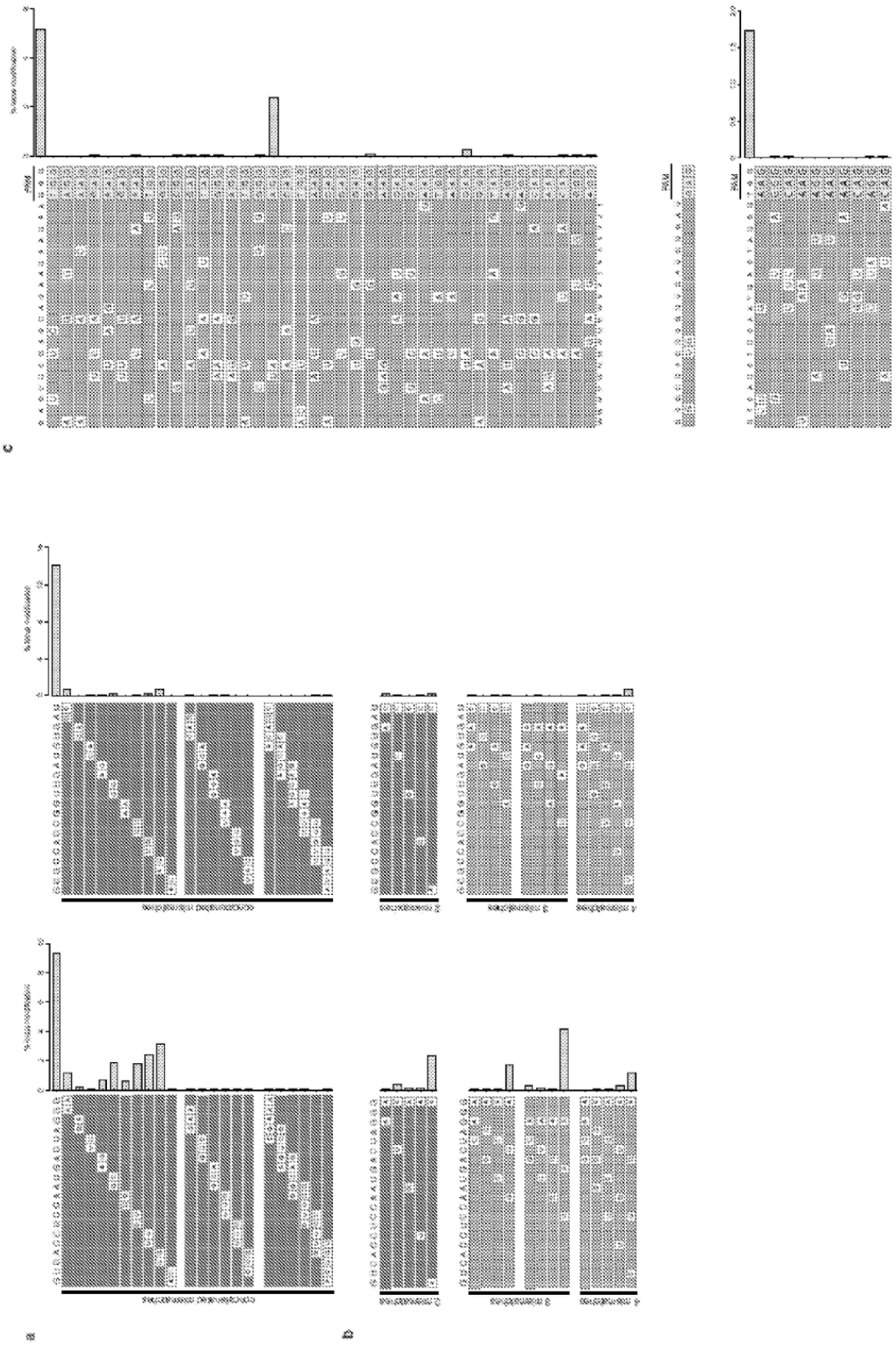


FIG. 13

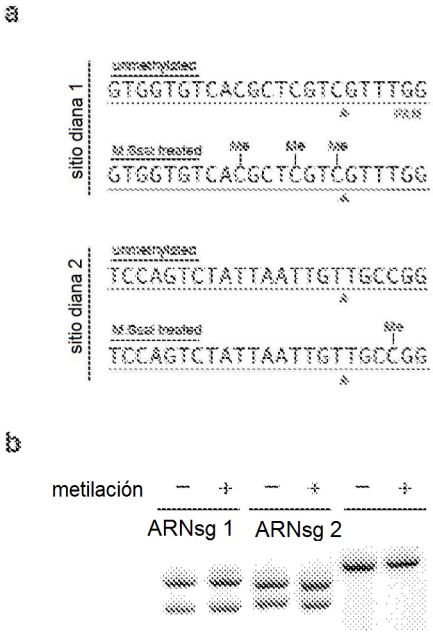

a
sitio diana 1
GTGGTGTCACGCTCGTCGTTTGG
M.SssI treated
GTGGTGTCACGCTCGTCGTTTGG
sitio diana 2
TCCAGTCTATTAATTGTTGCCGG
M.SssI treated
TCCAGTCTATTAATTGTTGCCGG
b
metilación
ARNsg 1
ARNsg 2

FIG. 14

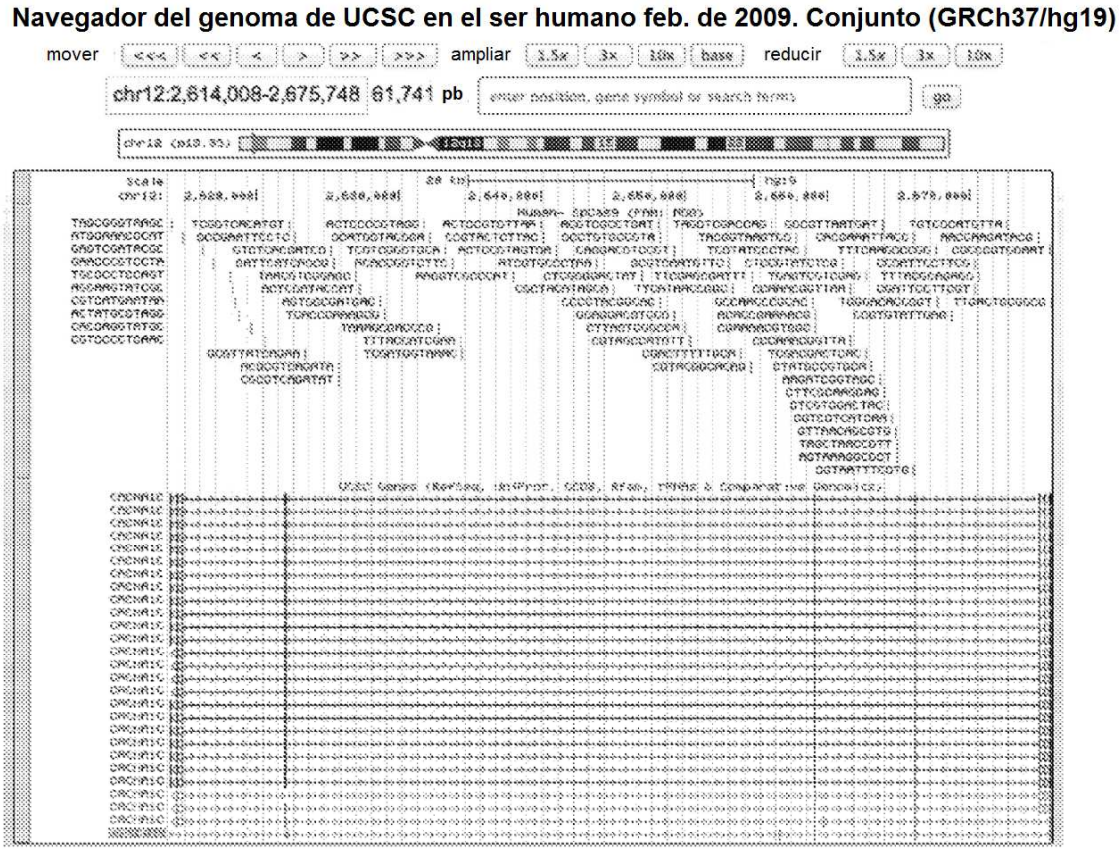

Navegador del genoma de UCSC en el ser humano feb. de 2009. Conjunto (GRCh37/hg19)
mover
ampliar
reducir
chr12:2,614,008-2,675,748
61,741 pb
go

FIG. 15

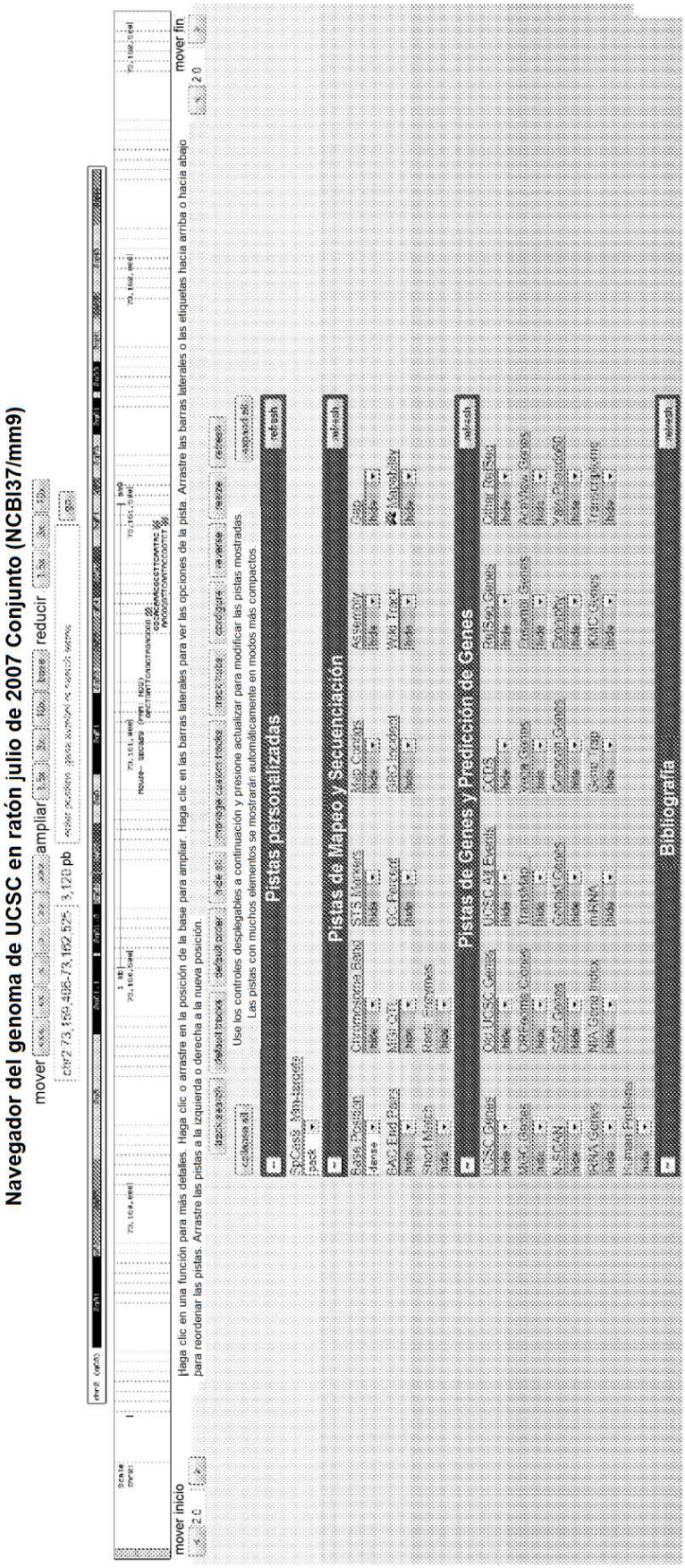

Navegador del genoma de UCSC en ratón julio de 2007 Conjunto (NCBI37/mm9)
mover
ampliar
reducir
chr2: 73,159,488-73,162,525
3,128 pb
mover inicio
mover fin
Haga clic en una función para más detalles. Haga clic o arrastre en la posición de la base para ampliar. Haga clic en las barras laterales para ver las opciones de la pista. Arrastre las barras laterales o las etiquetas hacia arriba o hacia abajo para reordenar las pistas. Arrastre las pistas a la izquierda o derecha a la nueva posición.
Pistas personalizadas
Pistas de Mapeo y Secuenciación
Pistas de Genes y Predicción de Genes
Bibliografía

FIG. 16

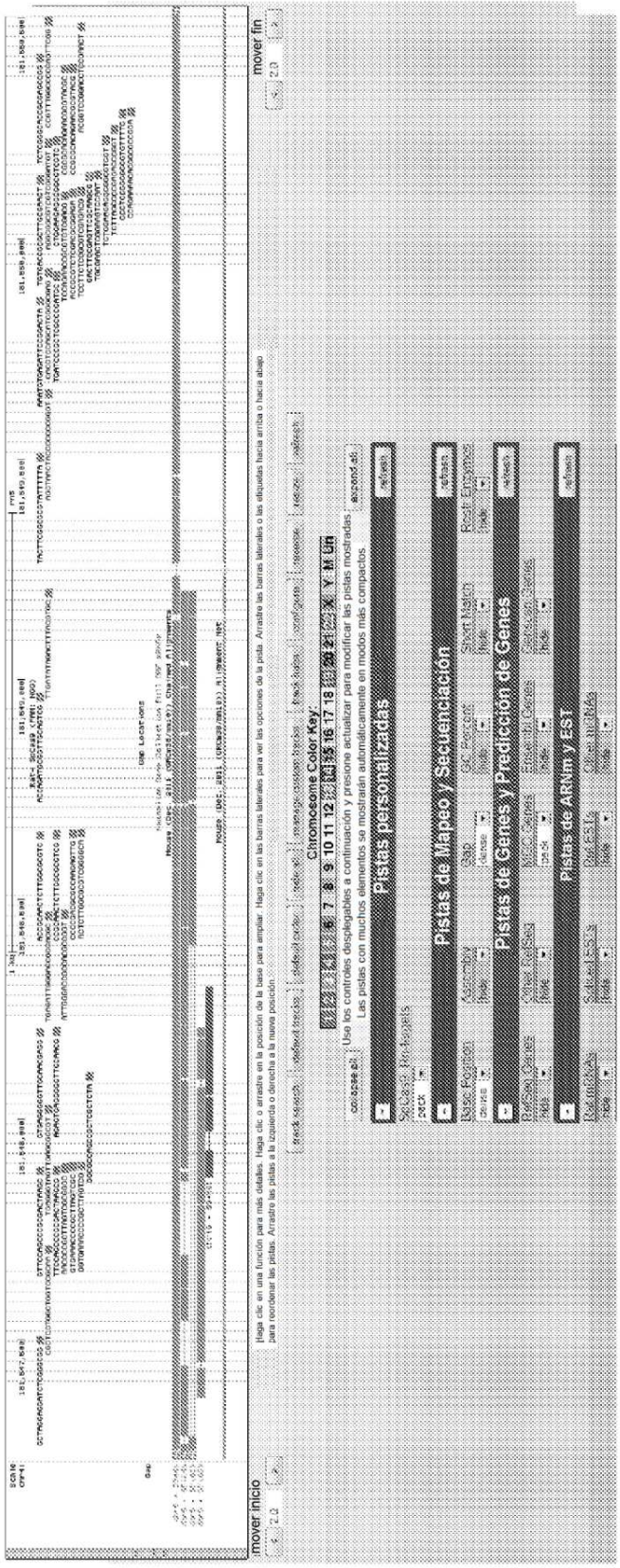

Navegador del genoma de UCSC en rata marzo de 2012 Conjunto (RGSC 5.0/rn5)
mover
ampliar
reducir
chr4:181,547,229-181,550,528 3,300 pb
ir
mover inicio
mover fin
Haga clic en una función para más detalles. Haga clic o arrastre en la posición de la base para ampliar. Haga clic en las barras laterales para ver las opciones de la pista. Arrastre las barras laterales o las etiquetas hacia arriba o hacia abajo para reordenar las pistas. Arrastre las pistas a la izquierda o derecha a la nueva posición.
Chromosome Color Key:
Use los controles desplegables a continuación y presione actualizar para modificar las pistas mostradas
Las pistas con muchos elementos se mostrarán automáticamente en modos más compactos
Pistas personalizadas
Pistas de Mapeo y Secuencia
Pistas de Genes y Predicción de Genes
Pistas de ARNm y EST
refrescar

FIG. 17

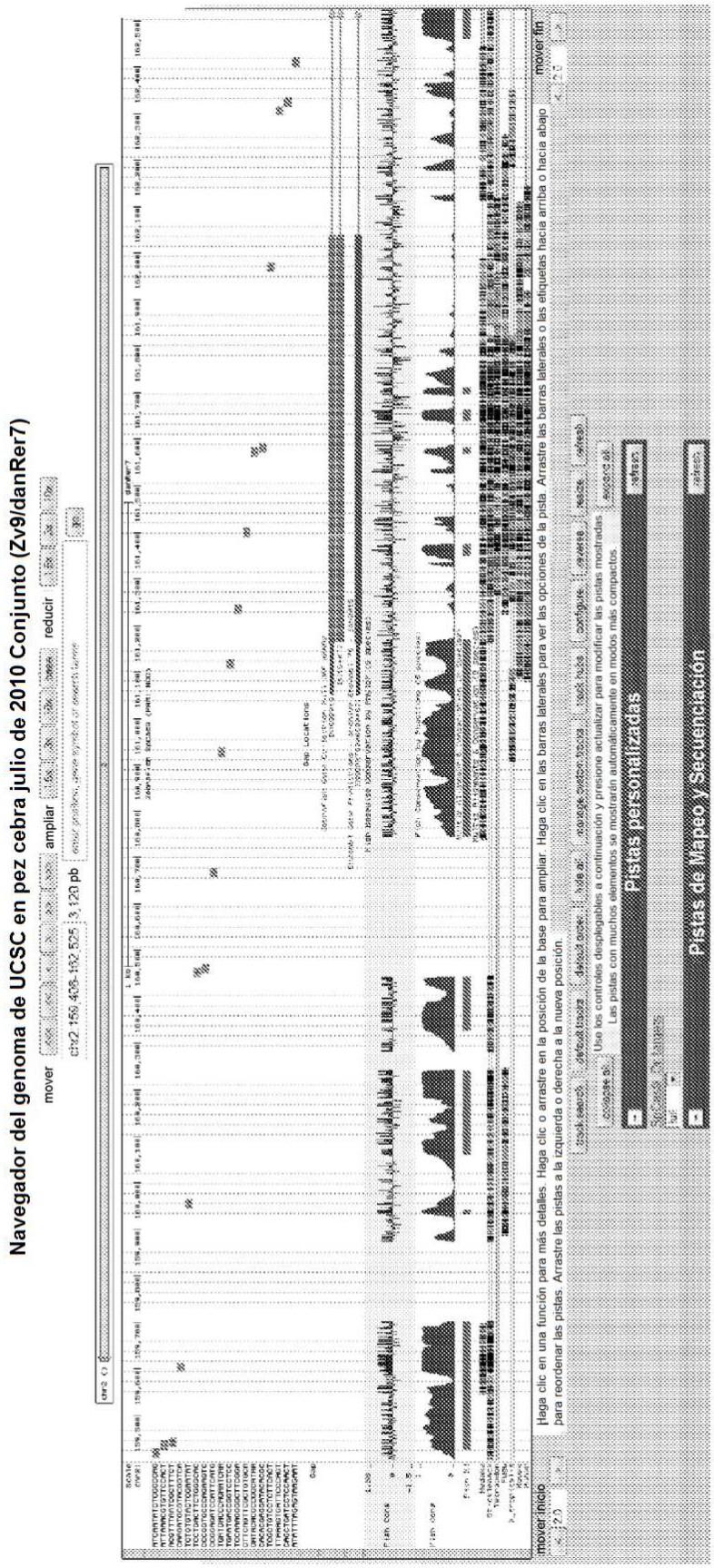

Navegador del genoma de UCSC en pez cebra julio de 2010 Conjunto (Zv9/danRer7)
mover
ampliar
reducir
chr2:159,408,152-525 | 3,120 pb
Pistas de Mapeo y Secuenciación
Pistas personalizadas
mover inicio
Haga clic en una función para más detalles. Haga clic o arrastre en la posición de la base para ampliar. Haga clic en las barras laterales para ver las opciones de la pista. Arrastre las barras laterales o las etiquetas hacia arriba o hacia abajo para reordenar las pistas. Arrastre las pistas a la izquierda o derecha a la nueva posición.
mover fin

FIG. 18

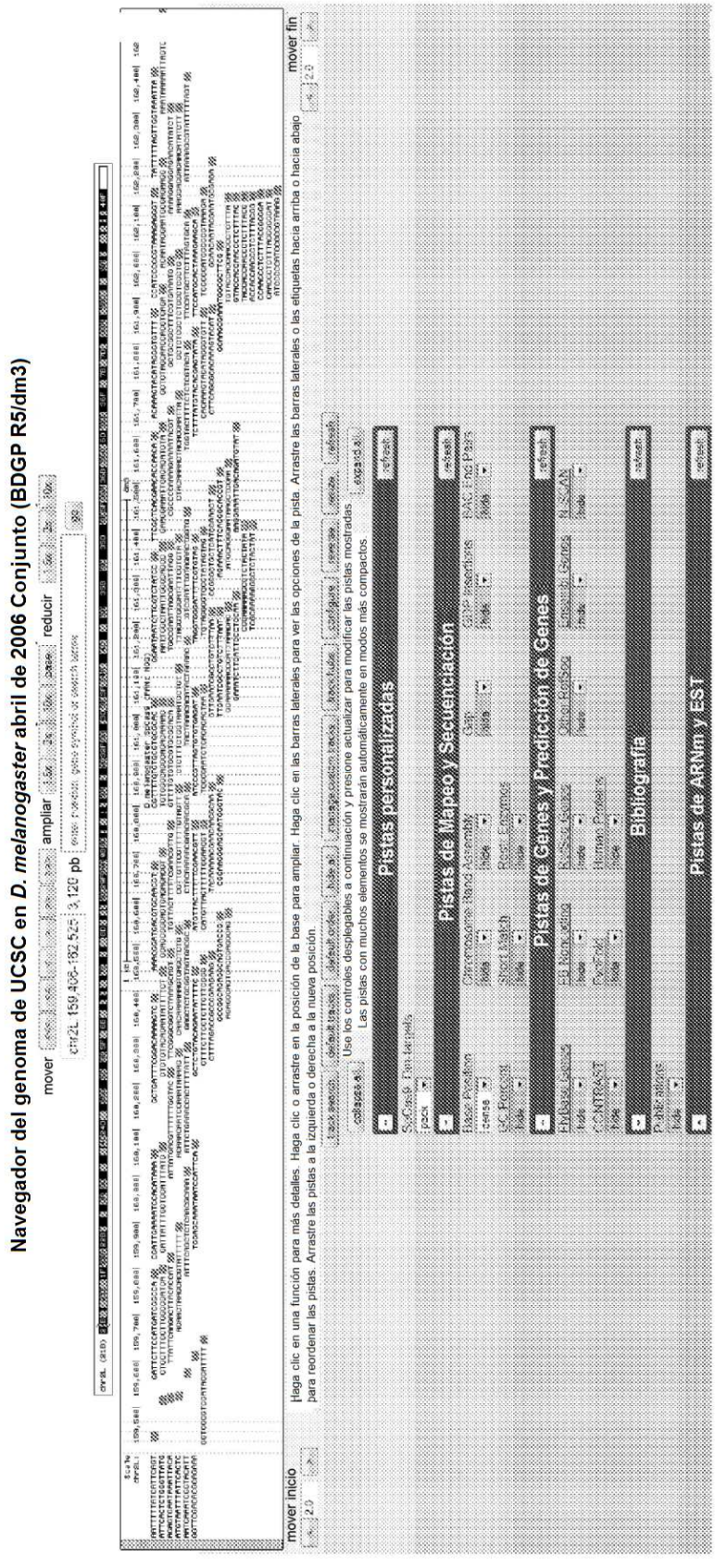

Navegador del genoma de UCSC en D. melanogaster abril de 2006 Conjunto (BDGP R5/dm3)
mover
ampliar
reducir
chr2L: 159,406-182,525; 3,120 pb
mover inicio
mover fin
Haga clic en una función para más detalles. Haga clic o arrastre en la posición de la base para ampliar. Haga clic en las barras laterales para ver las opciones de la pista. Arrastre las barras laterales o las etiquetas hacia arriba o hacia abajo para reordenar las pistas. Arrastre las pistas a la izquierda o derecha a la nueva posición.
Track search
default order
hide all
Use los controles desplegables a continuación y presione actualizar para modificar las pistas mostradas
Las pistas con muchos elementos se mostrarán automáticamente en modos más compactos
collapse all
expand all
Pistas personalizadas
Pistas de Mapeo y Secuenciación
Pistas de Genes y Predicción de Genes
Bibliografía
Pistas de ARNm y EST

FIG. 19

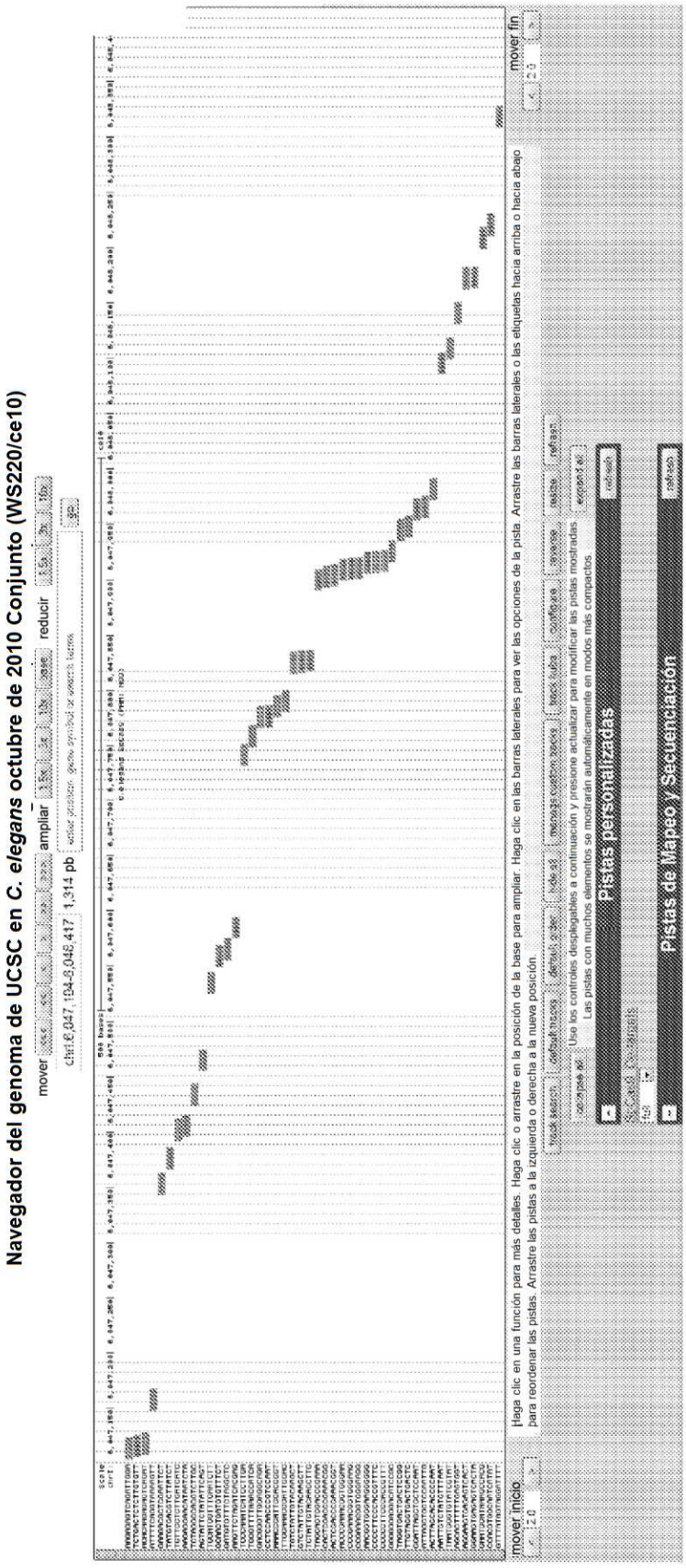

Navegador del genoma de UCSC en C. elegans octubre de 2010 Conjunto (WS220/ce10)
mover
ampliar
reducir
chrIII:6,047,104-6,046,417 1,314 pb
Pistas de Mapeo y Secuenciación
Pistas personalizadas
mover inicio
mover fin

FIG. 20

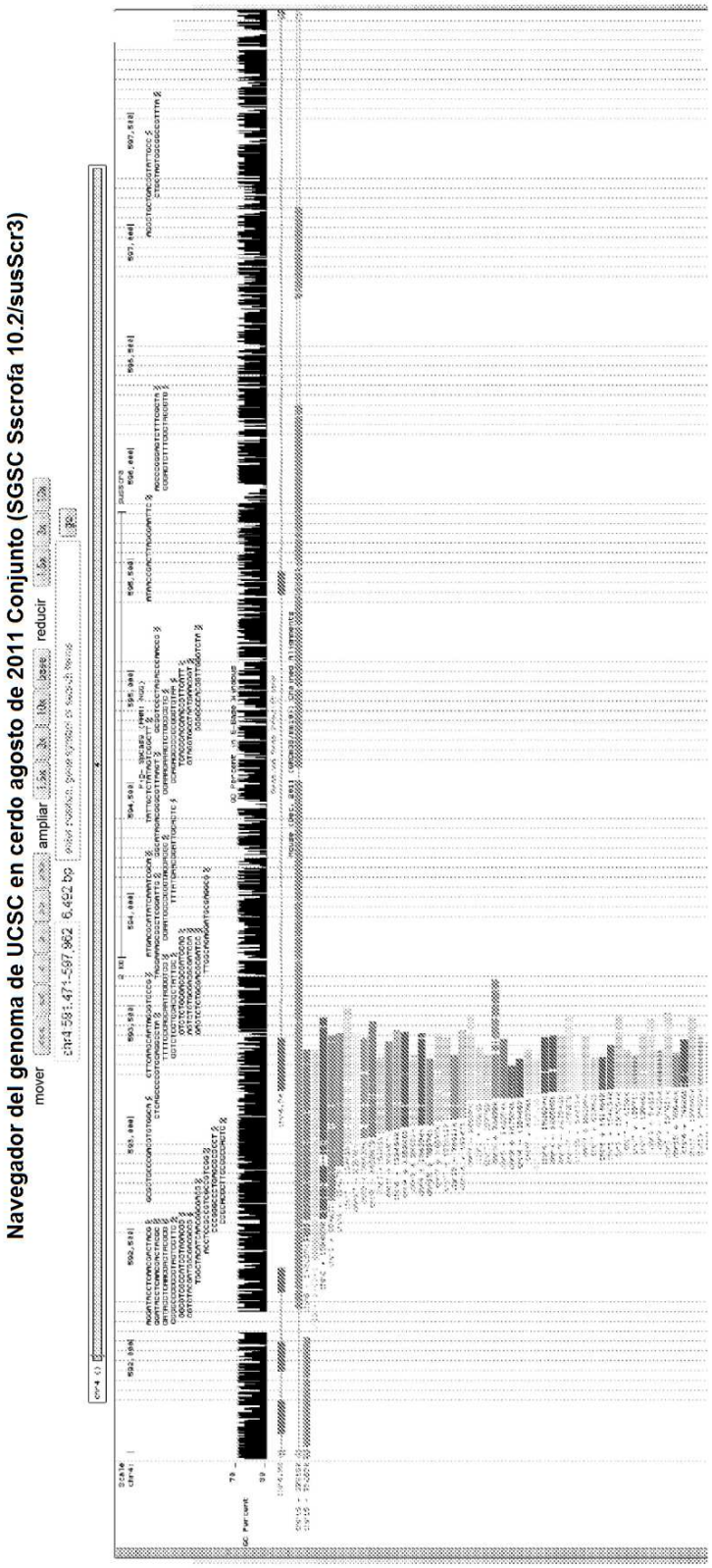

Navegador del genoma de UCSC en cerdo agosto de 2011 Conjunto (SGSC Sscrofa 10.2/susScr3)
mover
ampliar
reducir
chr4:581,471-587,962 6,492 bp

FIG. 21

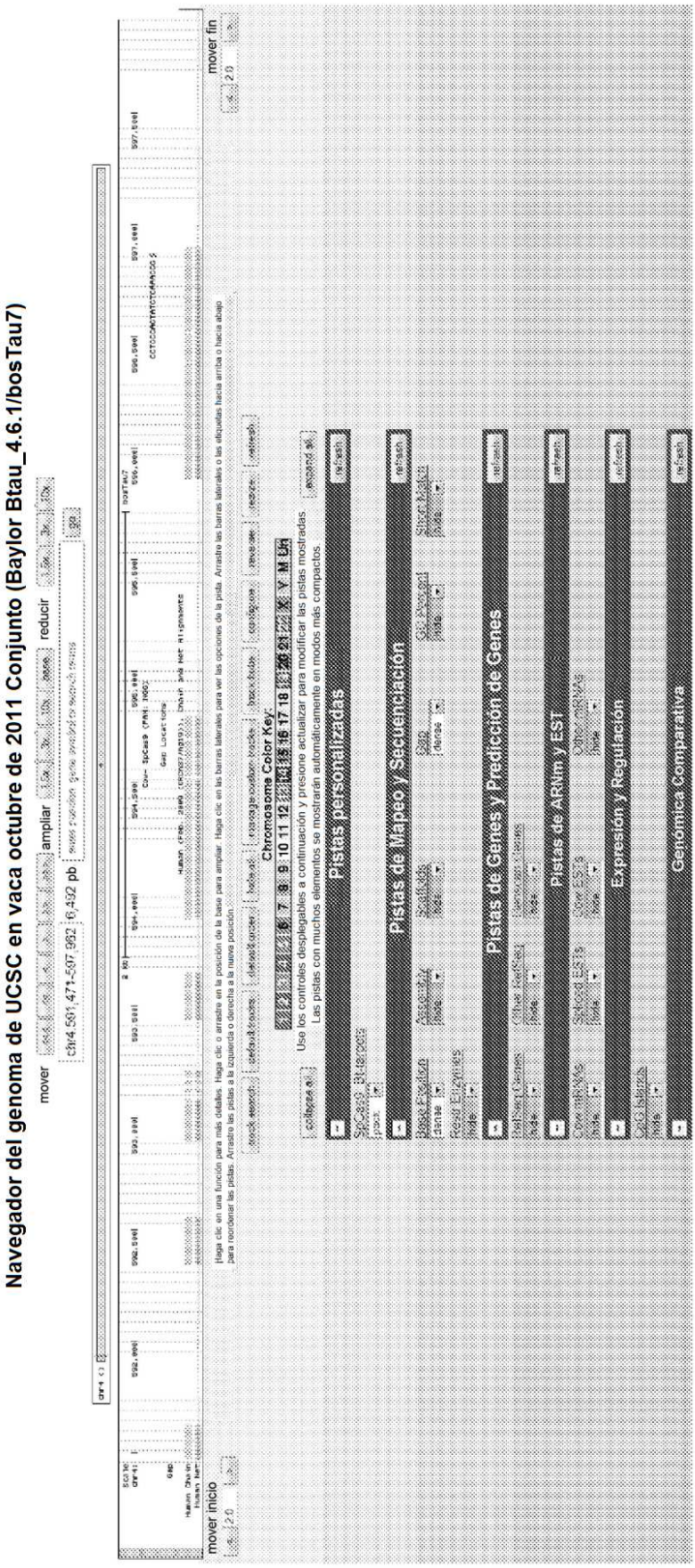

Navegador del genoma de UCSC en vaca octubre de 2011 Conjunto (Baylor Btau_4.6.1/bosTau7)
mover
ampliar
reducir
chr4:591,471-597,963 6,492 pb
mover inicio
mover fin
Chromosome Color Key:
Pistas personalizadas
Pistas de Mapeo y Secuenciación
Pistas de Genes y Predicción de Genes
Pistas de ARNm y EST
Expresión y Regulación
Genómica Comparativa

FIG. 22

Recurso: Herramienta Web de Diseño CRISPR

http://www.genome-engineering.org/tools

Entrada:

Secuencia de ADN de interés (23-500 pb de longitud)

Por ejemplo, exón 3 de *EMX1*: CGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGA ACCAGCCCAATGCGGACGACATCGATGTCACCTTCAATGACTAGCGTGGCCAACCACAAACCCACGAG

Salida:

Lista clasificada de todas las posibles secuencias de ARNsg, con detalles en las guías superiores

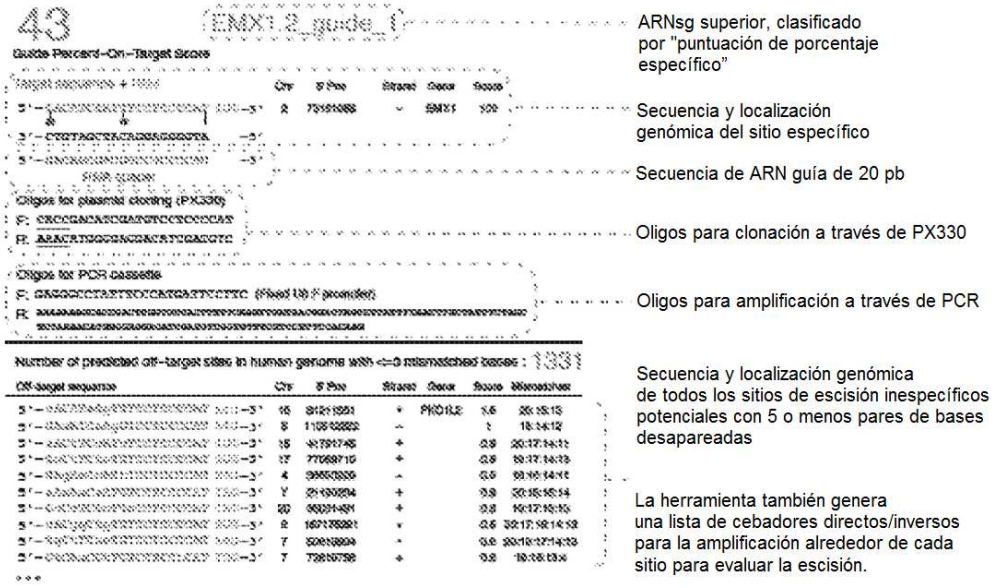


FIG. 23

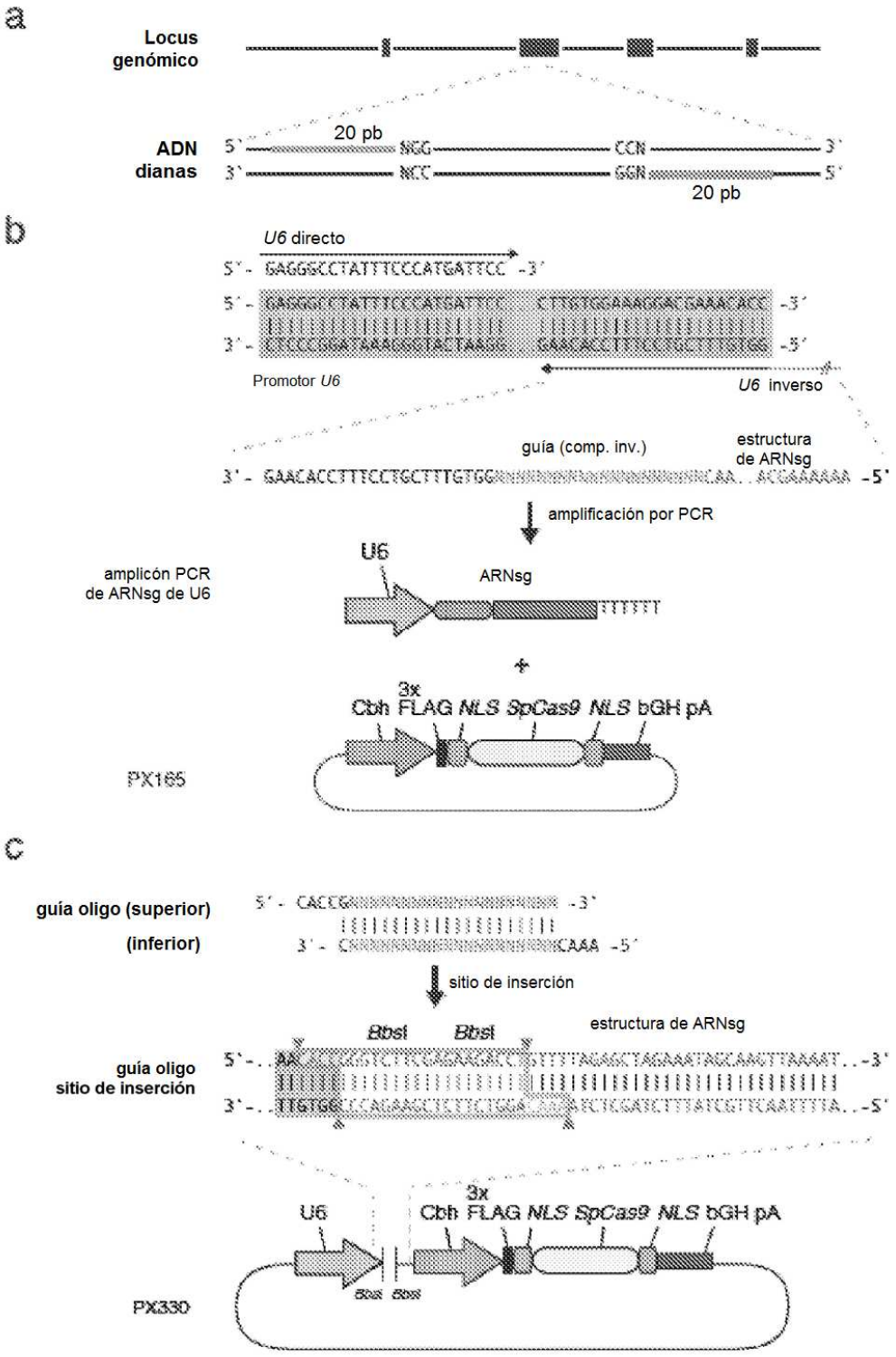

a
Locus genómico
ADN dianas
20 pb
NGG
CCN
NCC
GGN
20 pb
b
U6 directo
Promotor U6
U6 inverso
guía (comp. inv.)
estructura de ARNsg
amplificación por PCR
amplicón PCR de ARNsg de U6
U6
ARNsg
TTTTTT
3x
Cbh FLAG NLS SpCas9 NLS bGH pA
PX165
c
guía oligo (superior)
(inferior)
sitio de inserción
BbsI
BbsI
estructura de ARNsg
guía oligo sitio de inserción
U6
3x
Cbh FLAG NLS SpCas9 NLS bGH pA
PX330

FIG. 24

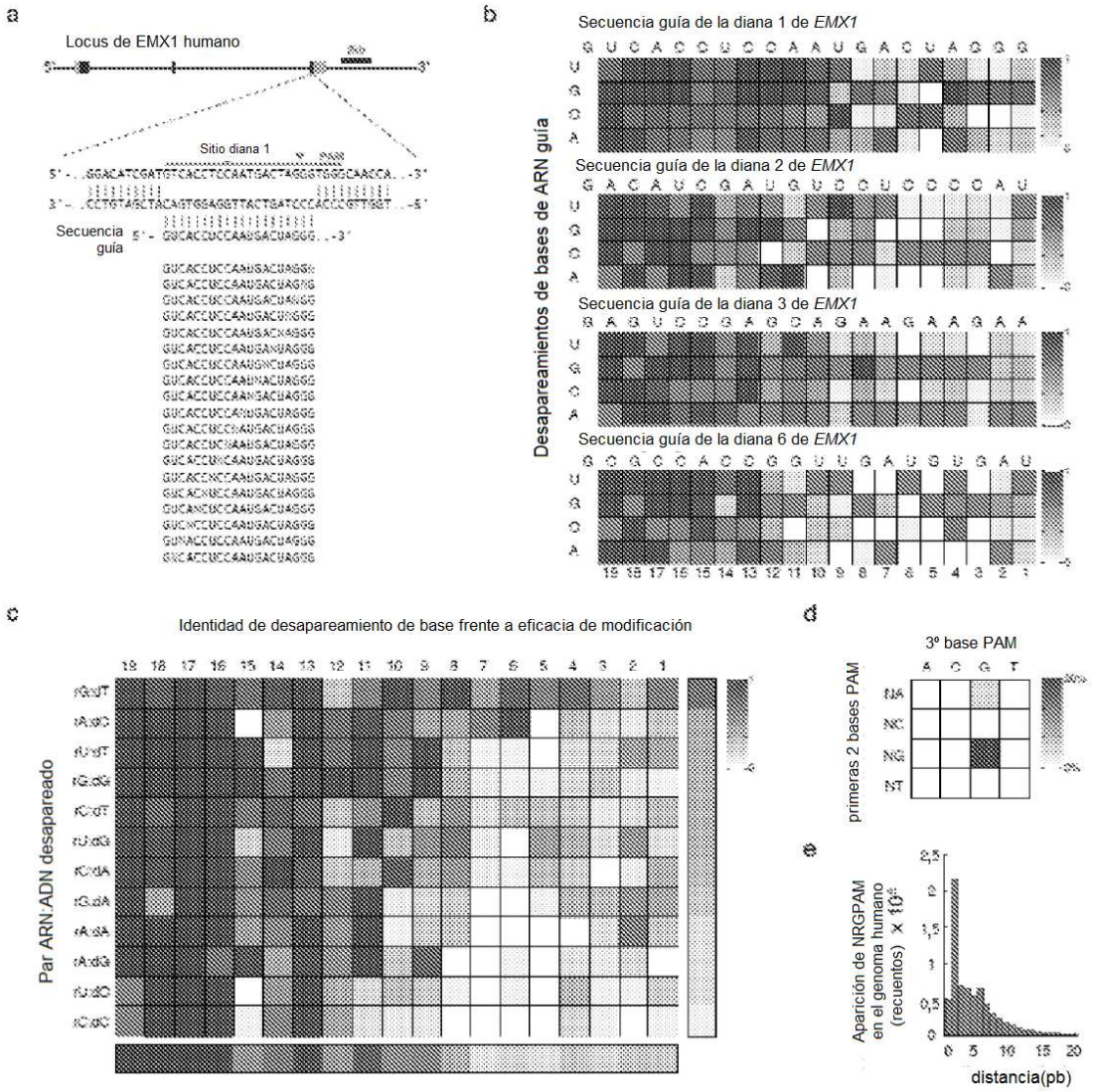

a
Locus de EMX1 humano
Sitio diana 1
Secuencia guía
b
Secuencia guía de la diana 1 de EMX1
Secuencia guía de la diana 2 de EMX1
Secuencia guía de la diana 3 de EMX1
Secuencia guía de la diana 6 de EMX1
Desapareamientos de bases de ARN guía
c
Identidad de desapareamiento de base frente a eficacia de modificación
Par ARN:ADN desapareado
d
3º base PAM
primeras 2 bases PAM
e
Aparición de NRGPAM en el genoma humano (recuentos) x 10^6
distancia(pb)

FIG. 25

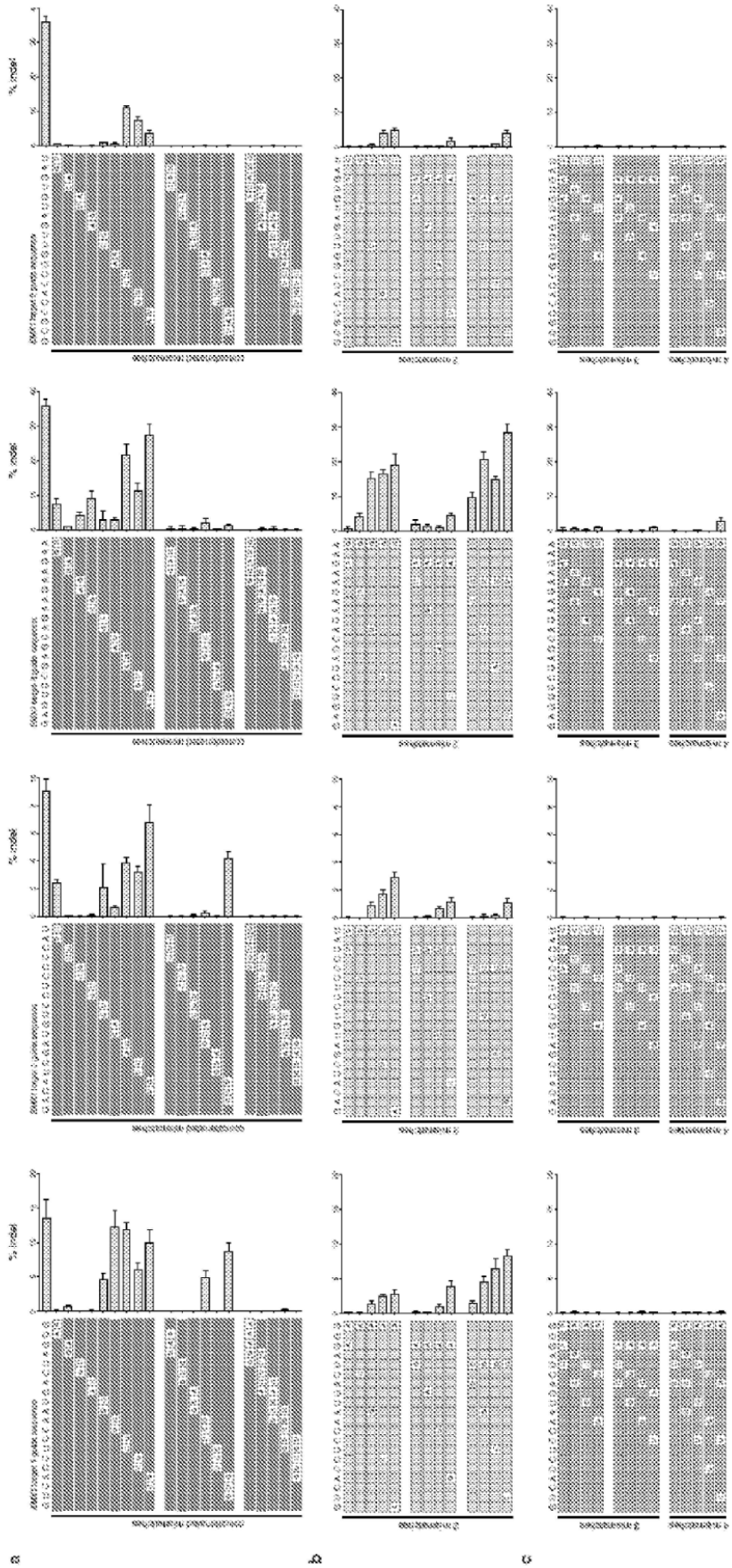


FIG. 26

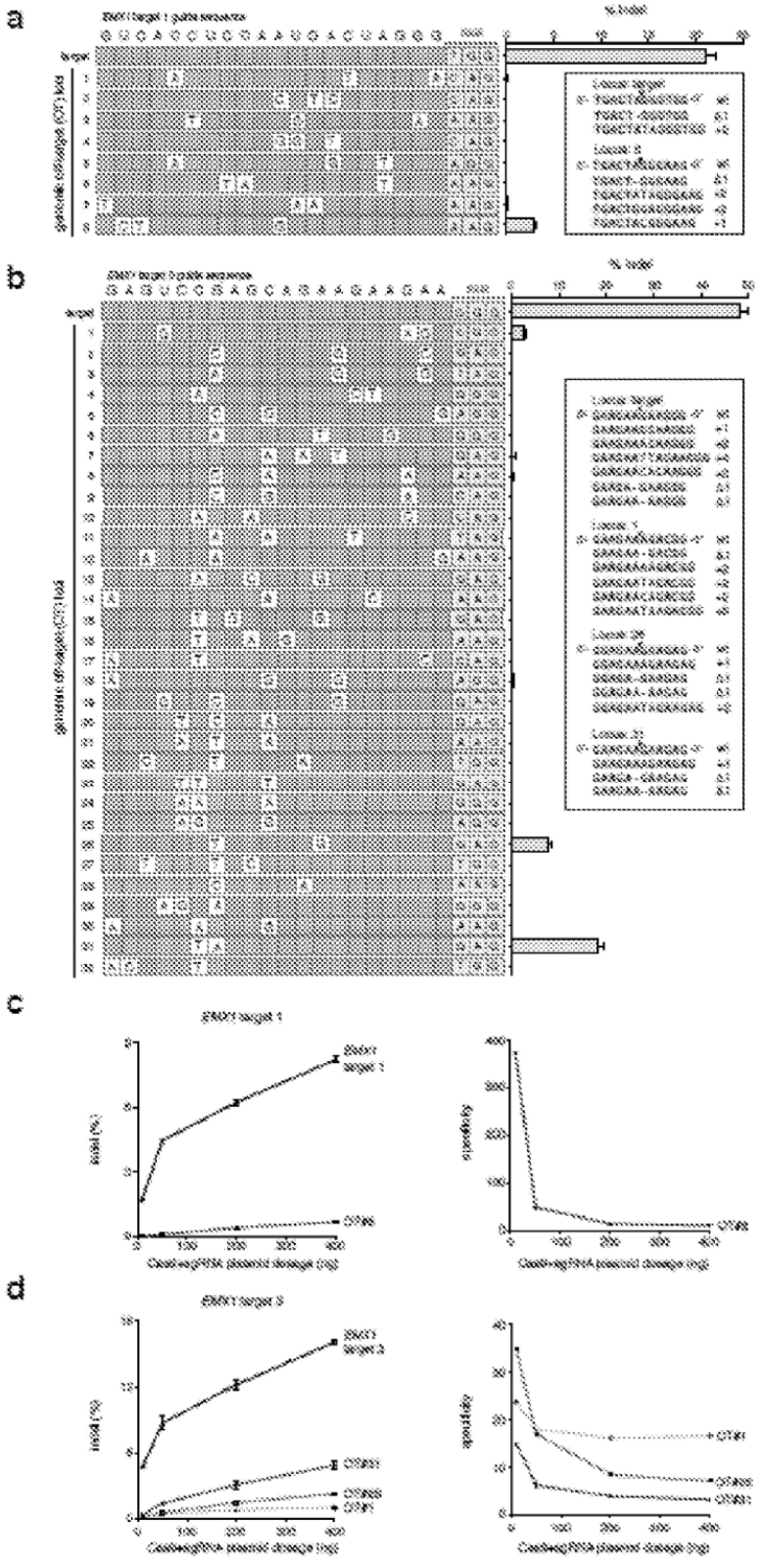
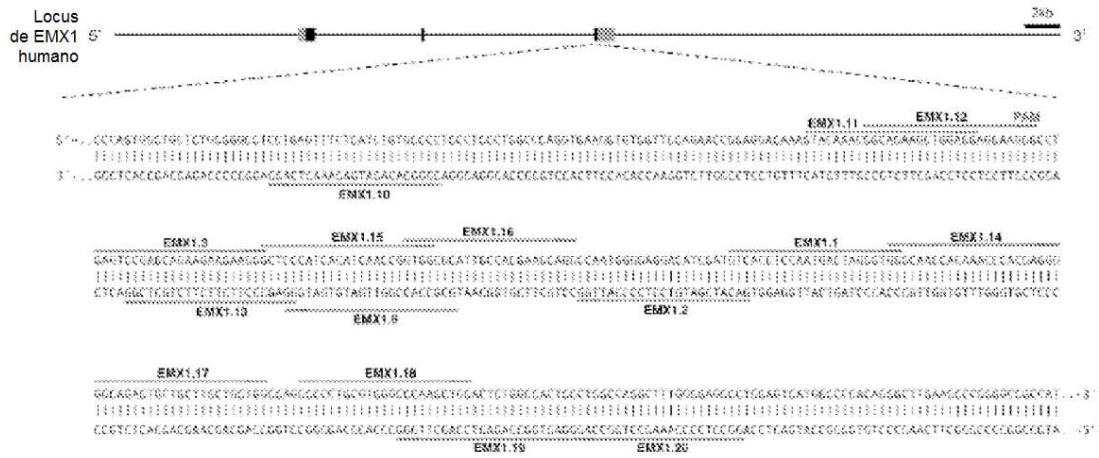


FIG. 27

a



b

Especies diana	Gen	Protoespaciador ID	Sitio diana (5´a 3´)	PAM	Cadena
Homo sapiens	*EMX1*	1	GTCACCTCCAATGACTAGGG	TGG	+
	EMX1	2	GACATCGATGTCCTCCCCAT	TGG	−
	EMX1	3	GAGTCCGAGCAGAAGAAGAA	GGG	+
	EMX1	6	GCGCCACCGGTTGATGTGAT	GGG	−
	EMX1	10	GGGGCACAGATGAGAAACTC	AGG	−
	EMX1	11	GTACAAACGGCAGAAGCTGG	AGG	+
	EMX1	12	GGCAGAAGCTGGAGGAGGAA	GGG	+
	EMX1	13	GGAGCCCTTCTTCTTCTGCT	CGG	−
	EMX1	14	GGGCAACCACAAACCCACGA	GGG	+
	EMX1	15	GCTCCCATCACATCAACCGG	TGG	+
	EMX1	16	GTGGCGCATTGCCACGAAGC	AGG	+
	EMX1	17	GGCAGAGTGCTGCTTGCTGC	TGG	+
	EMX1	18	GCCCCTGCGTGGGCCCAAGC	TGG	+
	EMX1	19	GAGTGGCCAGAGTCCAGCTT	GGG	−
	EMX1	20	GGCCTCCCCAAAGCCTGGCC	AGG	−

FIG. 28

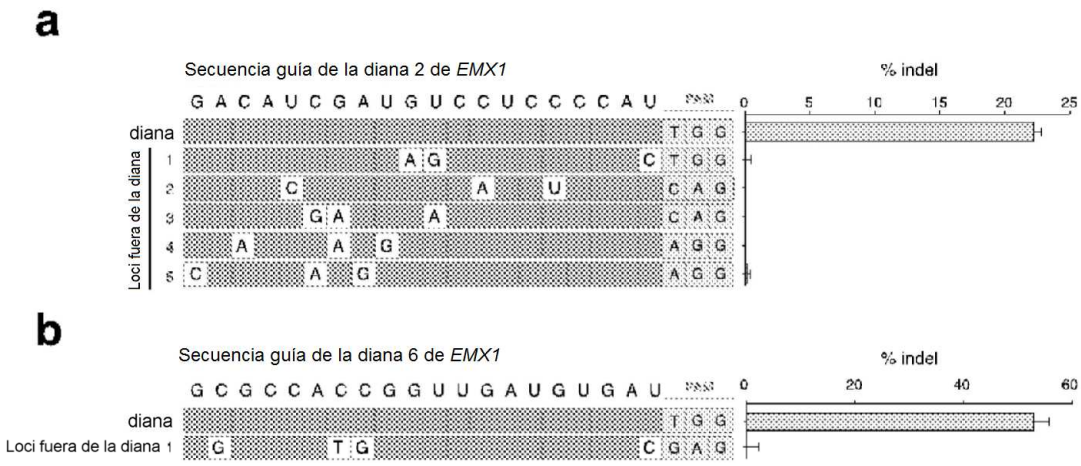

a
Secuencia guía de la diana 2 de EMX1
% indel
G A C A U C G A U G U C C U C C C C A U PAM
0
5
10
15
20
25
diana
T G G
Loci fuera de la diana
1
2
3
4
5
b
Secuencia guía de la diana 6 de EMX1
% indel
G C G C C A C C G G U U G A U G U G A U PAM
0
20
40
60
diana
T G G
Loci fuera de la diana 1

FIG. 29

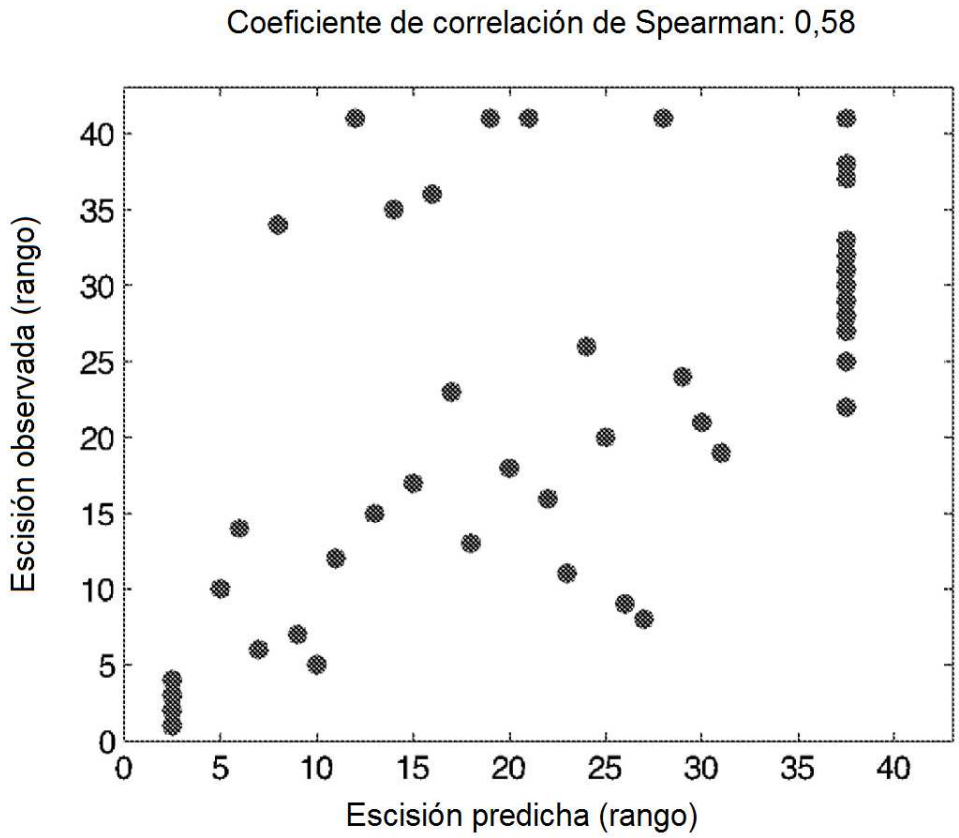
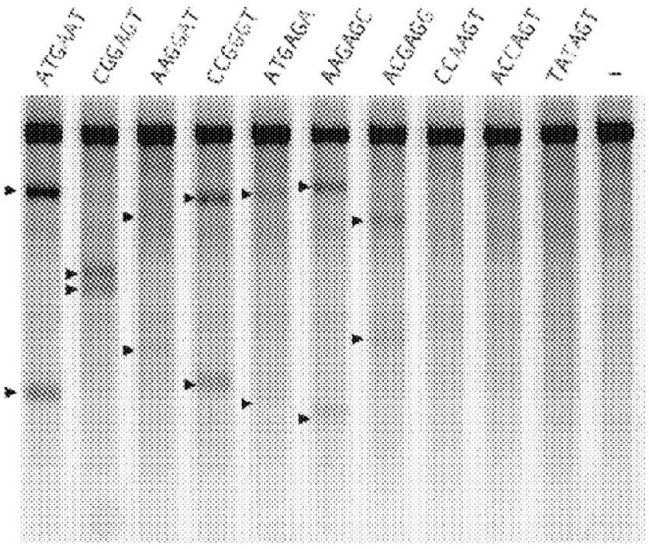

Coeficiente de correlación de Spearman: 0,58
Escisión observada (rango)
Escisión predicha (rango)
0
5
10
15
20
25
30
35
40

FIG. 30

Pruebas PAM: NNGRR



Secuencias espaciadoras para PAM probadas:

Espaciador	PAM
GCCCGGGTGGAACTGGTAGCC	ATGAAT
GTTGAAGATGAAGCCCAGAG	CGGAGT
GCTTCCGACGAGGTGGCCATC	AAGGAT
GCACCATCTCTCCGTGGTACC	CCGGGT
GGTGGAACTGGTAGCCATGA	ATGAGA
GCCATGAATGAGACCGACCCA	AAGAGC
GCATCCTCGTGGGCACTTCCG	ACGAGG
GCAGAGCGGAGTGCTGTTCTC	CCAAGT
GGTCGGTCTCATTCATGGCT	ACCAGT
GCAATAAAAGGTGCTATTGC	TATAGT

FIG. 31

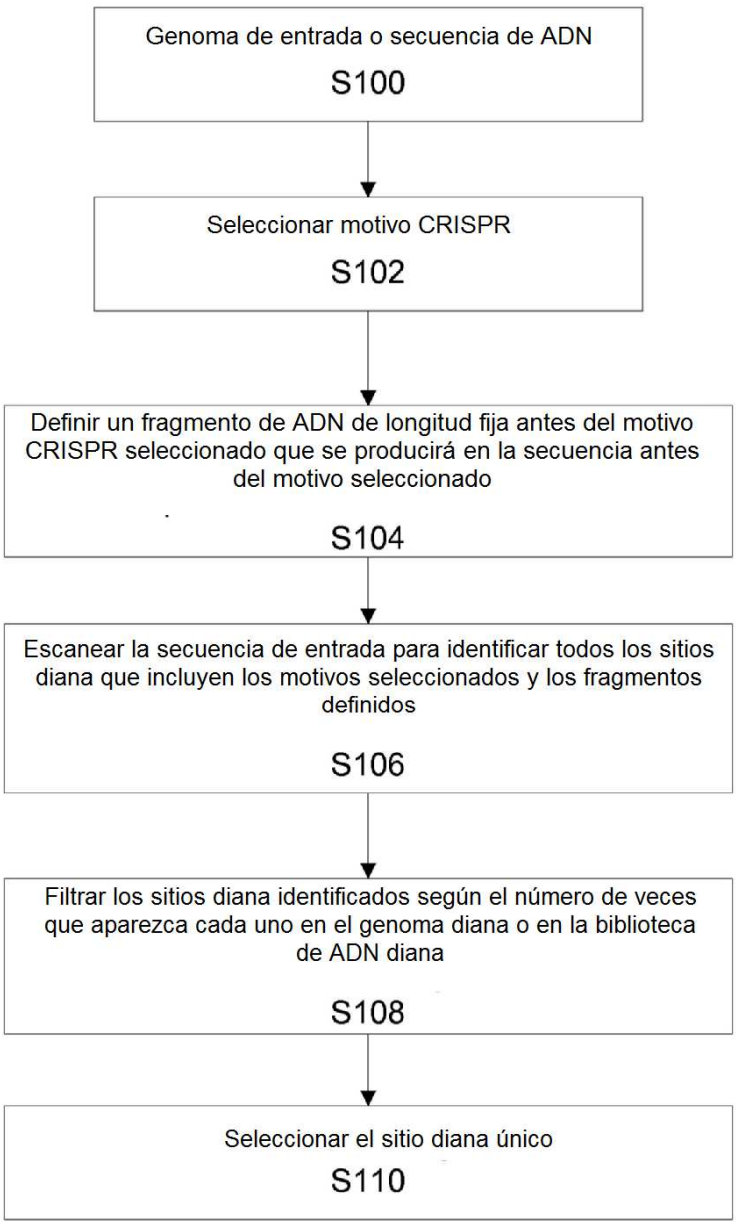

Genoma de entrada o secuencia de ADN
S100
Seleccionar motivo CRISPR
S102
Definir un fragmento de ADN de longitud fija antes del motivo CRISPR seleccionado que se producirá en la secuencia antes del motivo seleccionado
S104
Escanear la secuencia de entrada para identificar todos los sitios diana que incluyen los motivos seleccionados y los fragmentos definidos
S106
Filtrar los sitios diana identificados según el número de veces que aparezca cada uno en el genoma diana o en la biblioteca de ADN diana
S108
Seleccionar el sitio diana único
S110

FIG. 32

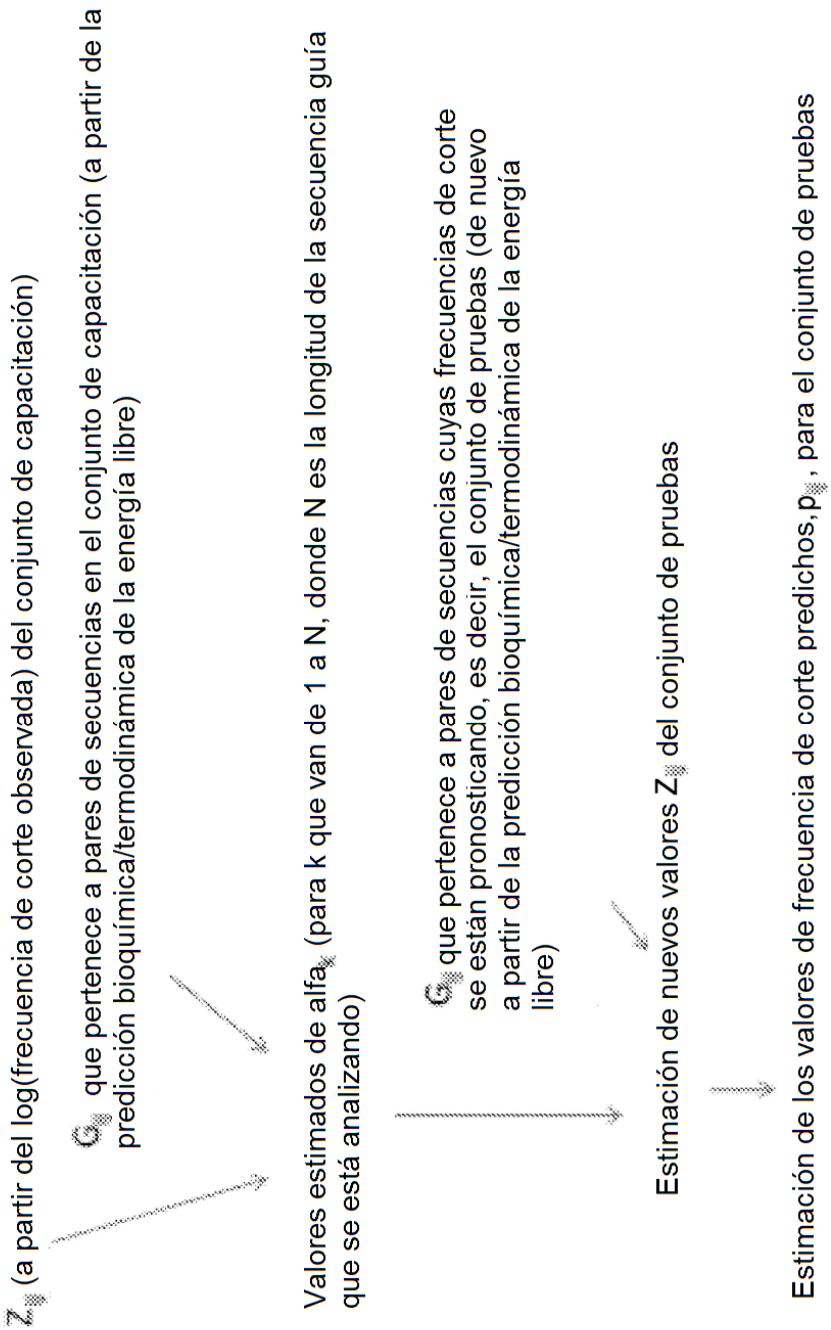

Z_{ij} (a partir del log(frecuencia de corte observada) del conjunto de capacitación)
G_{ij} que pertenece a pares de secuencias en el conjunto de capacitación (a partir de la predicción bioquímica/termodinámica de la energía libre)
Valores estimados de $alfa_k$ (para k que van de 1 a N, donde N es la longitud de la secuencia guía que se está analizando)
G_{ij} que pertenece a pares de secuencias cuyas frecuencias de corte se están pronosticando, es decir, el conjunto de pruebas (de nuevo a partir de la predicción bioquímica/termodinámica de la energía libre)
Estimación de nuevos valores Z_{ij} del conjunto de pruebas
Estimación de los valores de frecuencia de corte predichos, p_{ij}, para el conjunto de pruebas

FIG. 33A

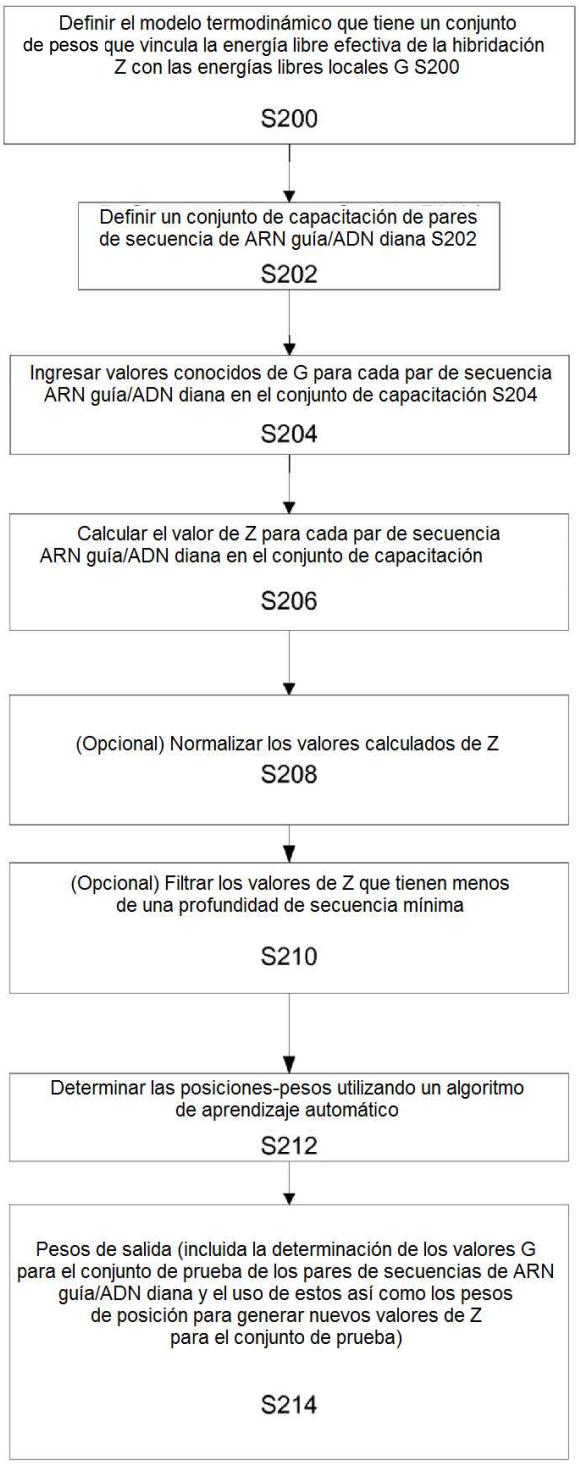

Definir el modelo termodinámico que tiene un conjunto de pesos que vincula la energía libre efectiva de la hibridación Z con las energías libres locales G S200
S200
Definir un conjunto de capacitación de pares de secuencia de ARN guía/ADN diana S202
S202
Ingresar valores conocidos de G para cada par de secuencia ARN guía/ADN diana en el conjunto de capacitación S204
S204
Calcular el valor de Z para cada par de secuencia ARN guía/ADN diana en el conjunto de capacitación
S206
(Opcional) Normalizar los valores calculados de Z
S208
(Opcional) Filtrar los valores de Z que tienen menos de una profundidad de secuencia mínima
S210
Determinar las posiciones-pesos utilizando un algoritmo de aprendizaje automático
S212
Pesos de salida (incluida la determinación de los valores G para el conjunto de prueba de los pares de secuencias de ARN guía/ADN diana y el uso de estos así como los pesos de posición para generar nuevos valores de Z para el conjunto de prueba)
S214

FIG. 33B

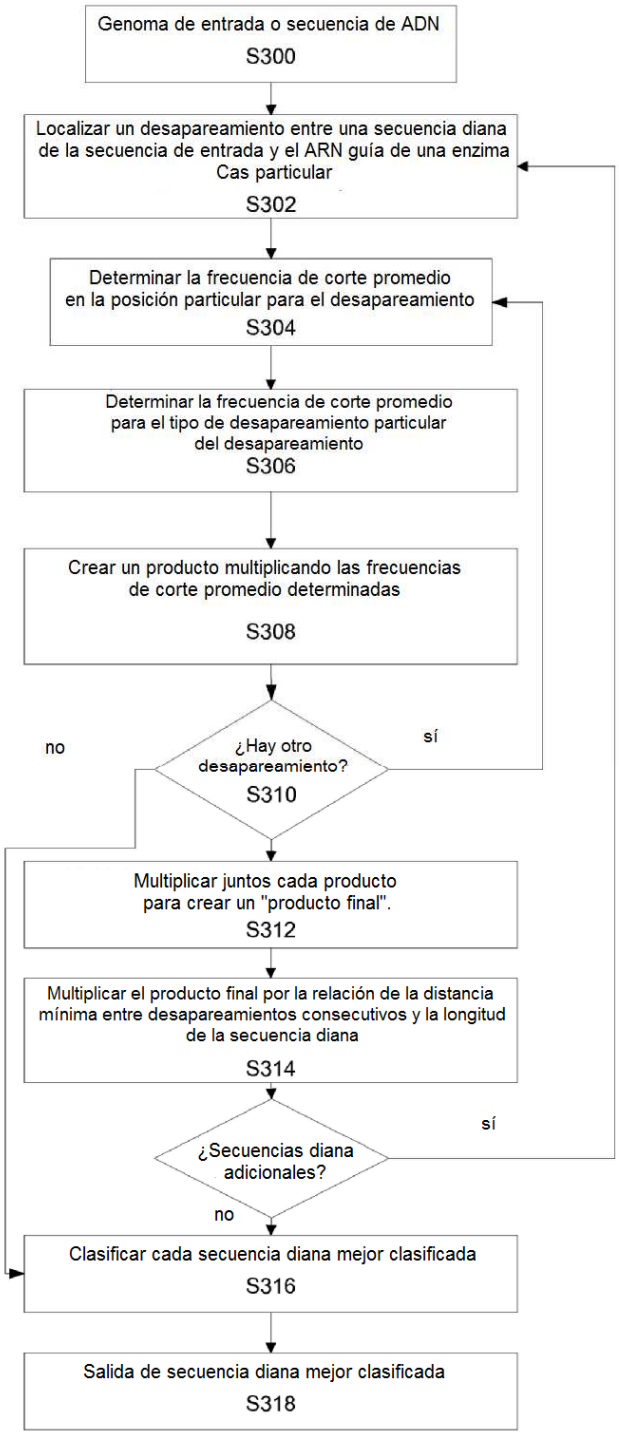

Genoma de entrada o secuencia de ADN
S300
Localizar un desapareamiento entre una secuencia diana de la secuencia de entrada y el ARN guía de una enzima Cas particular
S302
Determinar la frecuencia de corte promedio en la posición particular para el desapareamiento
S304
Determinar la frecuencia de corte promedio para el tipo de desapareamiento particular del desapareamiento
S306
Crear un producto multiplicando las frecuencias de corte promedio determinadas
S308
¿Hay otro desapareamiento?
S310
no
sí
Multiplicar juntos cada producto para crear un "producto final".
S312
Multiplicar el producto final por la relación de la distancia mínima entre desapareamientos consecutivos y la longitud de la secuencia diana
S314
¿Secuencias diana adicionales?
sí
no
Clasificar cada secuencia diana mejor clasificada
S316
Salida de secuencia diana mejor clasificada
S318

FIG. 34

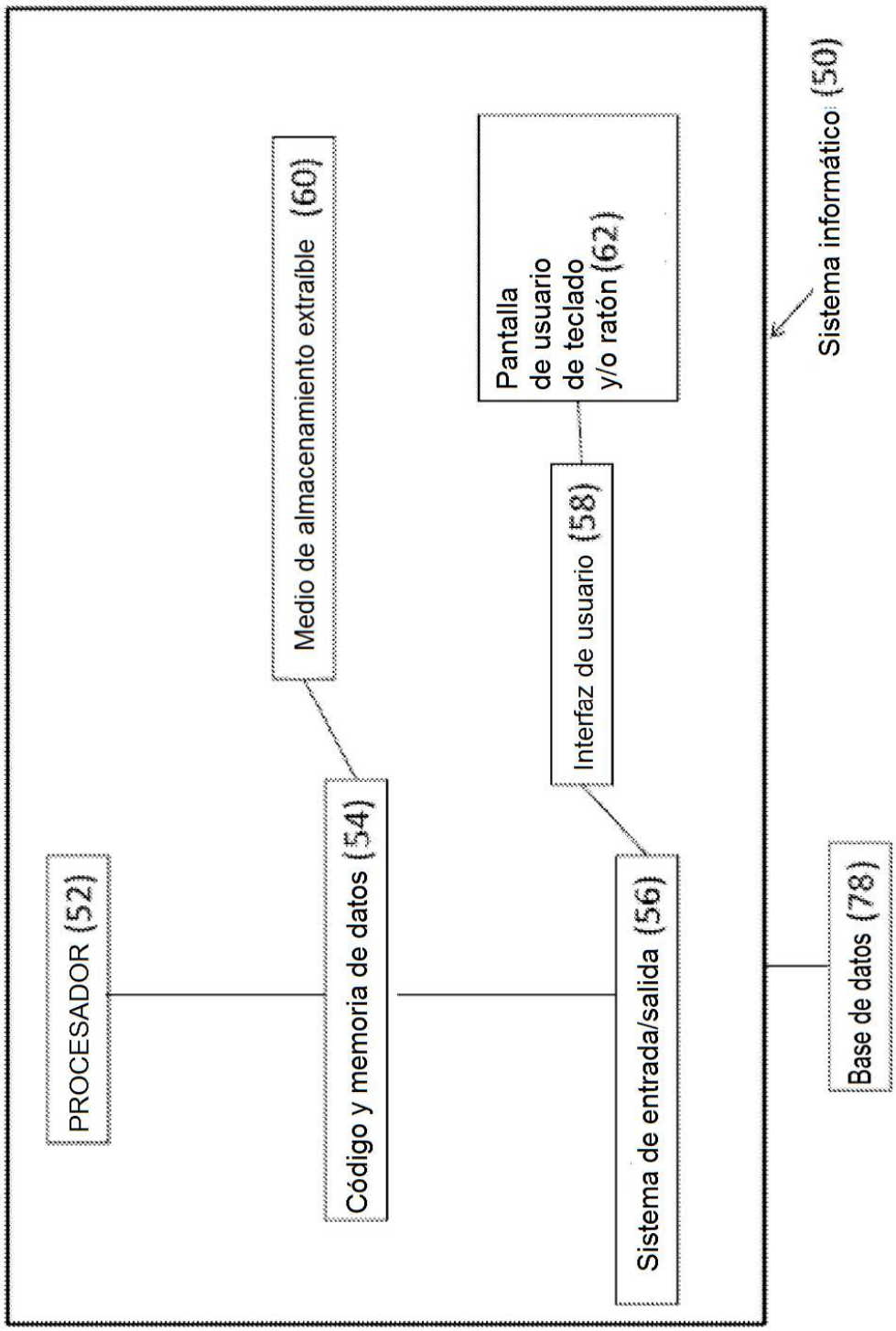

Sistema informático (50)
PROCESADOR (52)
Código y memoria de datos (54)
Medio de almacenamiento extraíble (60)
Sistema de entrada/salida (56)
Interfaz de usuario (58)
Pantalla de usuario de teclado y/o ratón (62)
Base de datos (78)

FIG. 35

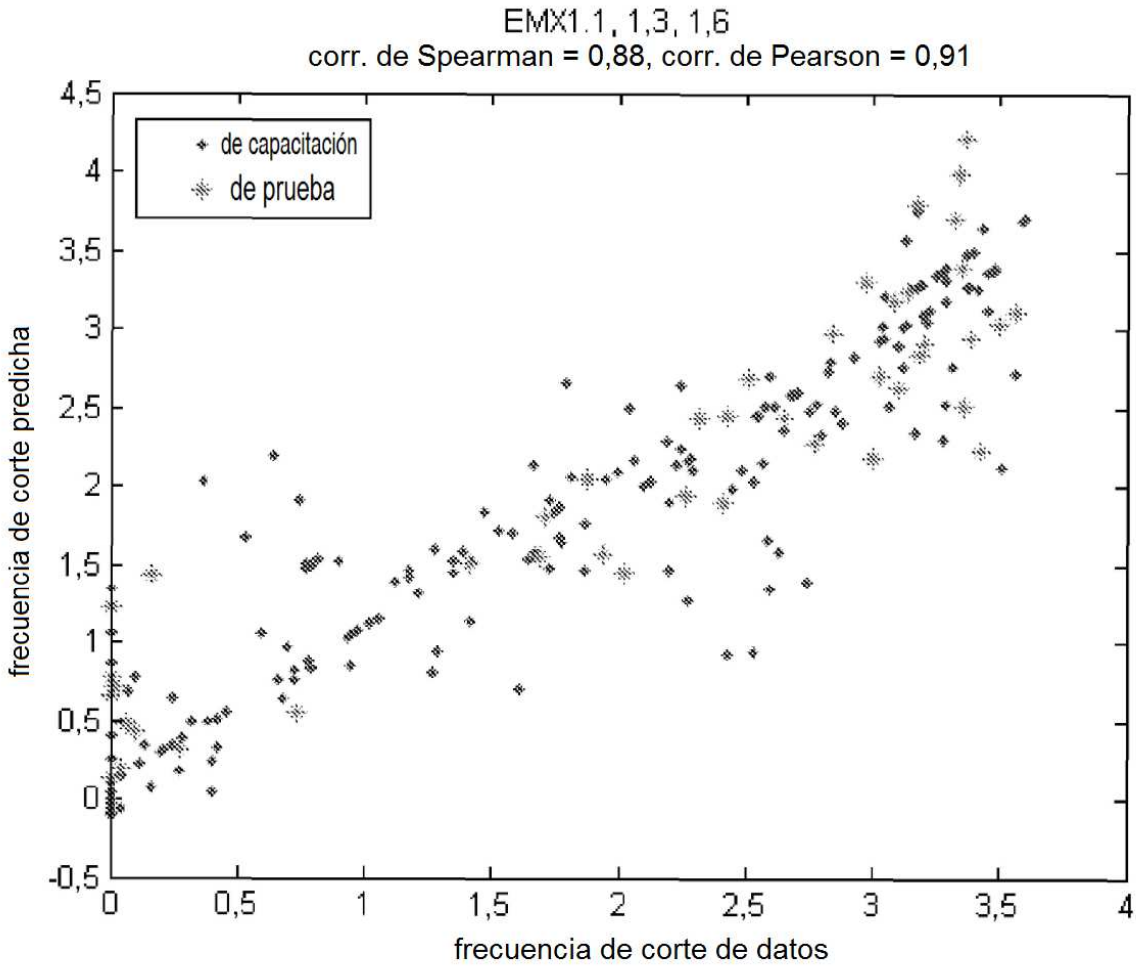

EMX1.1, 1,3, 1,6
corr. de Spearman = 0,88, corr. de Pearson = 0,91
de capacitación
de prueba
frecuencia de corte predicha
frecuencia de corte de datos
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
0
0,5
1
1,5
2
2,5
3
3,5
4

FIG. 36

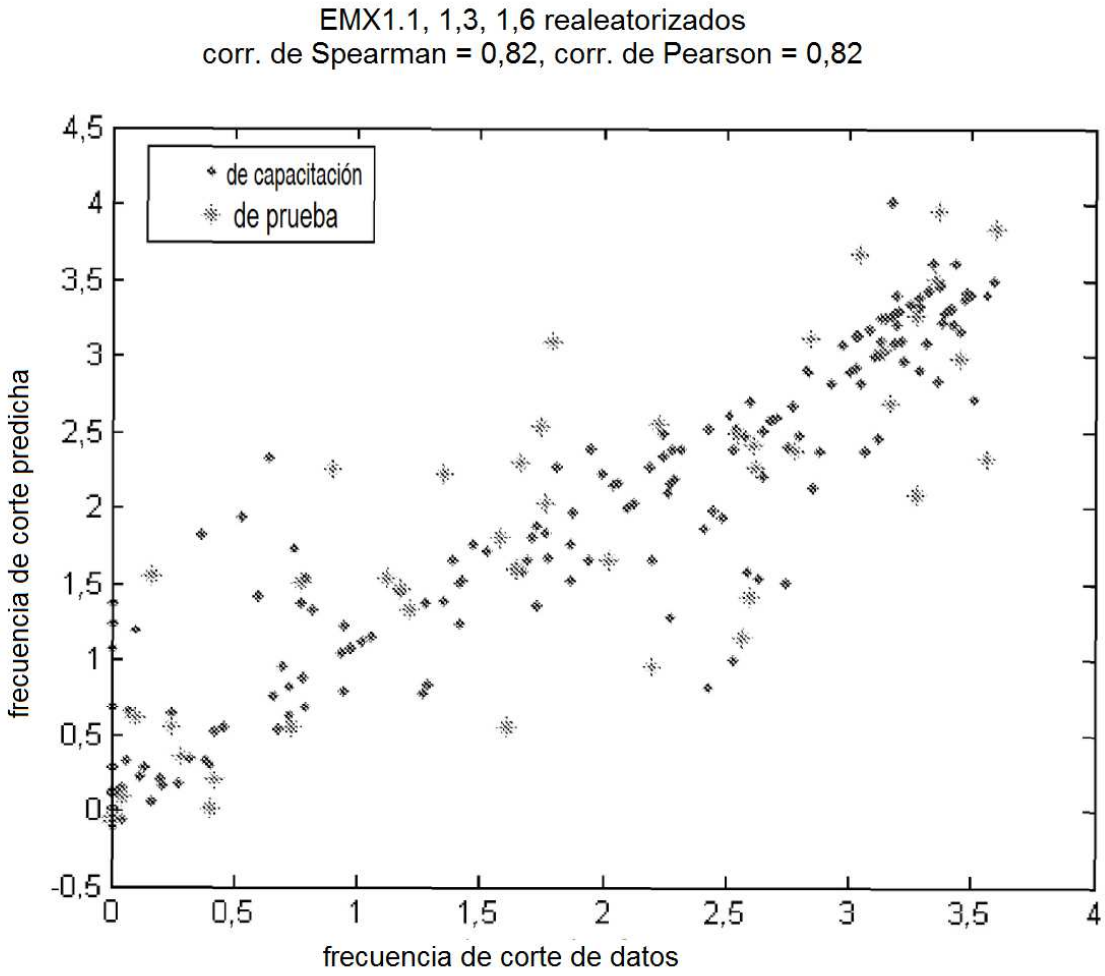

EMX1.1, 1,3, 1,6 realeatorizados
corr. de Spearman = 0,82, corr. de Pearson = 0,82
de capacitación
de prueba
frecuencia de corte predicha
frecuencia de corte de datos
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
0
0,5
1
1,5
2
2,5
3
3,5
4

FIG. 37