

19

OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA

11 Número de publicación: **2 970 263**

21 Número de solicitud: 202230911

51 Int. Cl.:

C07K 14/195 (2006.01)**C12N 15/63** (2006.01)

12

SOLICITUD DE PATENTE

A1

22 Fecha de presentación:

21.10.2022

43 Fecha de publicación de la solicitud:

27.05.2024

71 Solicitantes:

**UNIVERSITAT D'ALACANT / UNIVERSIDAD DE
ALICANTE (80.0%)****Carretera San Vicente del Raspeig s/n
03690 San Vicente del Raspeig (Alicante) ES;
CONSEJO SUPERIOR DE INVESTIGACIONES
CIENTÍFICAS (CSIC) (10.0%) y
CONSORCIO CENTRO DE INVESTIGACIÓN
BIOMÉDICA EN RED (10.0%)**

72 Inventor/es:

**ESQUERRA RUVIRA, Belén;
BAQUEDANO MOZOS, Ignacio;
RUIZ GONZÁLEZ, Raúl;
MARTÍNEZ MOJICA, Francisco Juan;
FERNÁNDEZ LÓPEZ, Almudena y
MONTOLIU JOSÉ, Lluís**

74 Agente/Representante:

ESCUADERO PRIETO, NicolásObservaciones:

La lista de secuencias está accesible al público en la página web de la OEPM para su descarga en formato electrónico.

54 Título: **PROTEÍNA ENDONUCLEASA Cas9 Y SISTEMA CRISPR-Cas ASOCIADO**

57 Resumen:

Proteína endonucleasa Cas9 y sistema CRISPR-Cas asociado.

La presente invención se refiere a una proteína endonucleasa Cas9 que comprende una secuencia aminoacídica según SEQ ID NO:1, a la secuencia nucleotídica que codifica dicha proteína, al vector de expresión que comprende dicha secuencia, a la célula que comprende dicha proteína, secuencia, y/o vector, al sistema CRISPR-Cas9 que comprende dicha proteína, al método para editar un genoma usando dicha proteína, y a los usos asociados de la misma para ingeniería genética y producción de antibacterianos.



FIG. 1

ES 2 970 263 A1

DESCRIPCIÓN

PROTEÍNA ENDONUCLEASA Cas9 Y SISTEMA CRISPR-Cas ASOCIADO

5 CAMPO DE LA INVENCION

La presente invención se enmarca en el campo de la ingeniería genética. Más concretamente, el objeto de la invención se refiere a nueva proteína endonucleasa Cas9 y a un sistema CRISPR-Cas que comprende dicha proteína para la edición genética en
10 células y producción de antibacterianos.

ANTECEDENTES DE LA INVENCION

Los sistemas CRISPR-Cas se han identificado en la mayoría de las arqueas y
15 aproximadamente la mitad de los genomas de bacterias como mecanismos de defensa contra la infección por ADN exógeno; esto es, plásmidos o virus (bacteriófagos). Cada sistema está constituido por una o varias agrupaciones de repeticiones de secuencias de ADN denominadas CRISPR (del inglés *C*lustered *R*egularly *I*nterspaced *S*hort *P*alindromic *R*epeats) y un conjunto de genes que codifican las proteínas Cas (del inglés *C*RISPR *a*ssociated), cuya presencia dará lugar a las endonucleasas encargadas de cortar y
20 degradar el ADN exógeno. Dichas repeticiones se encuentran regularmente espaciadas dentro de cada agrupación por secuencias no reiteradas denominadas 'espaciadores', al menos algunas de las cuales derivan de fragmentos genéticos de origen extracromosómico que el microorganismo adquiere tras entrar por primera vez en contacto
25 con un patógeno. Adyacente a cada agrupación de repeticiones-espaciadores hay una secuencia denominada 'líder' donde se localiza el promotor responsable de la transcripción de dicha agrupación en un ARN precursor (pre-crRNA) que abarca la totalidad de la misma. Este pre-crRNA es procesado mediante una ribonucleasa (RNAsa) dando lugar a crRNA maduros, cada uno de los cuales contiene un único espaciador. Cada uno de estos
30 crRNA, comúnmente denominado 'ARN guía', aparea con su secuencia complementaria en la molécula de ADN exógeno, activando con ello a una proteína Cas específica que produce la degradación de dicho ADN, protegiendo así a la célula huésped de la infección.

Más allá de su función como sistema natural de inmunidad adaptativa en organismos
35 procariotas, los sistemas CRISPR-Cas se han convertido en unas de las herramientas de edición genética más poderosas en los campos de la biología, la biomedicina y la

biotecnología. En concreto, permiten el silenciamiento o la eliminación de genes, mutagénesis, y correcciones de secuencias específicas del genoma de cualquier célula de una manera fácil, rápida, y altamente precisa [Jian, W. et al. Nat. Biotechnol., 2013, 31 (3), 233-239; Mali, P. et al. Science, 2013, 339 (6121), 823-826]. Entre sus numerosas aplicaciones destacan el diagnóstico y tratamiento de enfermedades [Srivastava, S., Upadhyay, D. J., & Srivastava, A. Front. Mol. Biosci., 2020, 7, 378; Jolany vangah, S. et al. Biol Proced Online, 2020, 22 (1), 1-14] y la producción de antimicrobianos específicos de secuencia [Bikard, D. et al. Nat. Biotechnol., 2014, 32 (11), 1146-1150].

- 10 Los sistemas CRISPR-Cas de Clase 2 - Tipo II (también denominados sistemas CRISPR-Cas9) son los más utilizados como herramienta de edición genética debido a la alta tasa de eficiencia de la endonucleasa Cas9. Además, a diferencia de los sistemas de Clase 1, el ARN guía (gRNA) consiste en dos moléculas de ARN parcialmente apareadas entre sí, formando un híbrido tracrRNA:crRNA, que comprende la secuencia activadora tracrRNA,
- 15 un pequeño ARN no codificante con dos funciones críticas: disparar el procesamiento del pre-crRNA por la enzima RNasa III y, subsecuentemente, servir como nexo de unión entre el crRNA y Cas9 para dirigirla hacia la secuencia diana del ADN bicatenario que debe degradar. Bajo esta configuración, el complejo Cas9:crRNA:tracrRNA escanea dicho ADN en busca de una secuencia corta (1-10 nucleótidos) llamada PAM (del inglés, *Protospacer*
- 20 *Adjacent Motif*), que se encuentra 3-4 nucleótidos aguas abajo del sitio de corte de Cas9. Cuando el dominio PI de Cas9 reconoce dicha secuencia PAM, el ADN bicatenario se desestabiliza y ocurre el apareamiento de bases entre el ADN y el crRNA, dando lugar al heterodúplex tracrRNA:crRNA:ADN, de aproximadamente 20 pares de bases, que se posicionará dentro de Cas9 en el surco central entre los lóbulos REC y NUC. Una vez
- 25 formado este complejo cuaternario (tracrRNA:crRNA:ADN y Cas9) el dominio HNH del lóbulo NUC se aproximará a la cadena complementaria de la secuencia diana provocando su escisión y lo mismo ocurrirá con el dominio RuvC con la cadena no complementaria. Como consecuencia, entre ambos dominios tendrá lugar un corte de doble cadena (DSB, por sus siglas en inglés). Como alternativa a la guía dual nativa tracrRNA:crRNA, para
- 30 guiar a las proteínas Cas9 se puede utilizar una molécula de ARN (sgRNA) [Jinek, M. et al. (2012). Science, 337(6096), 816-821], que combina parte de las secuencias del crRNA y el tracrRNA.

De entre todos los sistemas CRISPR-Cas9 destacan los basados en la proteína Cas9 de la bacteria *Streptococcus pyogenes* (SpCas9), que requiere la presencia de una secuencia PAM excepcionalmente corta (5'-NGG-3') para el reconocimiento de la secuencia diana,

lo cual supone una gran ventaja respecto a otras proteínas Cas9. Sin embargo, su gran tamaño supone una limitación para su administración, especialmente en ensayos *in vivo* con células eucariotas. Se requiere, por tanto, la identificación y caracterización bioquímica y funcional de proteínas Cas9 alternativas de menor tamaño.

5

La presente invención está orientada a resolver la limitación expuesta anteriormente mediante una nueva proteína endonucleasa Cas9 de pequeño tamaño (~120 kDa), apta para su uso en diversas herramientas de biología molecular para ingeniería genética equivalentes a las implementadas con otras endonucleasas Cas9, así como para la

10

DESCRIPCIÓN BREVE DE LA INVENCION

La presente invención soluciona el problema del estado de la técnica expuesto en la sección anterior al proporcionar una proteína endonucleasa Cas9 con un tamaño tal que facilita su administración tanto a células procariotas como eucariotas mediante vectores comúnmente empleados en biotecnología y biomedicina; esto es, plásmidos o bacteriófagos, en el caso de las bacterias, y virus adenoasociados (AAV, por sus siglas en inglés), para células de mamíferos. Además, a diferencia de las proteínas Cas9 del estado de la técnica, permite que se puedan incorporar, en tan solo una molécula vector (especialmente en el caso de los AAV), secuencias de elementos genéticos accesorios, como secuencias reguladoras o moldes para la edición genética.

15

20

Así pues, en un primer aspecto, la presente invención se refiere a una proteína endonucleasa Cas9 que comprende una secuencia aminoacídica según SEQ ID NO: 1 (de aquí en adelante, "proteína de la presente invención").

25

En una realización preferente, la proteína de la invención comprende una secuencia aminoacídica con al menos un 70% de identidad de secuencia con SEQ ID NO: 1. Concretamente, proteínas con una secuencia aminoacídica con al menos un 70, 75, 80, 85, 90, 95 y 100% de identidad de secuencia con SEQ ID NO: 1.

30

En el ámbito de interpretación de la presente invención, el término "identidad de secuencia" se entenderá como el grado de similitud entre dos secuencias nucleotídicas o aminoacídicas, expresado a modo de porcentaje, que se obtiene al alinear dichas secuencias. Este dependerá del número de nucleótidos o residuos comunes entre las

35

secuencias alineadas. Se determina mediante programas bioinformáticos bien establecidos en el estado de la técnica, tales como BLAST (del inglés, *Basic Local Alignment Search Tool*) o FASTA.

- 5 Se considerarán también dentro de la presente invención aquellas secuencias análogas, derivadas o equivalentes a SEQ ID NO: 1 que comprenden al menos un residuo de aminoácido alterado por una inserción, sustitución, delección, o modificación química de un aminoácido respecto a la secuencia aminoacídica de la proteína de la presente invención.
- 10 En una realización preferente, la proteína de la presente invención comprende una secuencia aminoacídica según SEQ ID NO: 5. En una realización aún más preferente, la secuencia aminoacídica comprende una inserción de al menos un aminoácido. En el ámbito de interpretación de la presente invención, se entenderá por “inserción” cualquier tipo de mutación en la secuencia aminoacídica de la proteína de la presente invención que
- 15 implique la adición de uno o más aminoácidos. La secuencia aminoacídica identificada como SEQ ID NO: 5 comprende una inserción de 19 aminoácidos tras el primer aminoácido de la secuencia identificada como SEQ ID NO: 1.

- En otra realización preferente, la proteína de la presente invención comprende una
- 20 secuencia aminoacídica según SEQ ID NO: 7. En una realización aún más preferente, la secuencia aminoacídica comprende una sustitución de al menos un aminoácido y una inserción de al menos un aminoácido. En el ámbito de interpretación de la presente invención, se entenderá por “sustitución” cualquier tipo de mutación en la secuencia aminoacídica de la proteína de la presente invención que implique el reemplazo de uno o
- 25 más aminoácidos. La secuencia aminoacídica identificada como SEQ ID NO: 7 comprende la mutación T2A y una inserción de 11 aminoácidos al final de la secuencia identificada como SEQ ID NO: 1.

- En un segundo aspecto, la presente invención se refiere a una secuencia nucleotídica que
- 30 codifica la proteína de la presente invención (de aquí en adelante, “secuencia nucleotídica de la presente invención”). En el ámbito de interpretación de la presente invención, se entenderá por “secuencia nucleotídica que codifica la proteína de la presente invención” cualquier secuencia de nucleótidos que, bajo un control de expresión adecuado, sea capaz de transcribir y traducir la secuencia aminoacídica de la proteína de la presente invención.

35

En una realización preferente, la secuencia nucleotídica de la presente invención comprende la secuencia nucleotídica identificada como SEQ ID NO: 2.

5 En otra realización preferente, la secuencia nucleotídica de la presente invención comprende la secuencia nucleotídica identificada como SEQ ID NO: 4.

En otra realización preferente, la secuencia nucleotídica de la presente invención comprende la secuencia nucleotídica identificada como SEQ ID NO: 6.

10 En otra realización preferente, la secuencia nucleotídica de la presente invención comprende la secuencia nucleotídica identificada como SEQ ID NO: 8.

En un tercer aspecto, la presente invención se refiere a un vector de expresión que comprende la secuencia nucleotídica de la presente invención (de aquí en adelante,
15 “vector de expresión de la presente invención”). En el ámbito de interpretación de la presente invención, se entenderá por “vector de expresión” cualquier molécula de ADN que pueda utilizarse como vehículo para transportar la secuencia nucleotídica de la presente invención al interior de una célula huésped. El vector de expresión de la presente invención puede comprender una molécula de ácido nucleico monocatenaria, bicatenaria
20 o parcialmente bicatenaria; una molécula de ADN, ARN, o híbrida ADN:ARN. Ejemplos de vectores de expresión son los plásmidos y los bacteriófagos o fagos.

En una realización preferente, el vector de expresión de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 2.
25

En otra realización preferente, el vector de expresión de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 4.

En otra realización preferente, el vector de expresión de la presente invención comprende
30 una secuencia nucleotídica según SEQ ID NO: 6.

En otra realización preferente, el vector de expresión de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 8.

35 En un cuarto aspecto, la presente invención se refiere a una célula que comprende la proteína de la presente invención, y/o la secuencia nucleotídica de la presente invención,

y/o el vector de expresión de la presente invención (de aquí en adelante, "célula de la presente invención"). En el ámbito de interpretación de la presente invención, se entenderá por "célula" cualquier unidad básica, estructural y funcional de un ser vivo susceptible a una o más de las siguientes alteraciones genéticas: transformación (absorción directa, incorporación y expresión de la secuencia nucleotídica de la presente invención), transfección o transducción (introducción de material genético externo mediante el vector de expresión de la invención), y translocación (introducción de la proteína de la presente invención en el interior del ribosoma).

- 10 En una realización preferente, la célula de la presente invención comprende una proteína con una secuencia aminoacídica con al menos un 70% de identidad de secuencia con SEQ ID NO: 1.

15 En otra realización preferente, la célula de la presente invención comprende una proteína con una secuencia aminoacídica según SEQ ID NO: 5.

En otra realización preferente, la célula de la presente invención comprende una proteína con una secuencia aminoacídica según SEQ ID NO: 7.

- 20 En otra realización preferente, la célula de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 2.

En otra realización preferente, la célula de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 4.

25

En otra realización preferente, la célula de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 6.

- 30 En otra realización preferente, la célula de la presente invención comprende una secuencia nucleotídica según SEQ ID NO: 8.

En otra realización preferente, la célula de la presente invención comprende un vector de expresión que comprende a su vez una secuencia nucleotídica según SEQ ID NO: 2.

- 35 En otra realización preferente, la célula de la presente invención comprende un vector de expresión que comprende a su vez una secuencia nucleotídica según SEQ ID NO: 4.

En otra realización preferente, la célula de la presente invención comprende un vector de expresión que comprende a su vez una secuencia nucleotídica según SEQ ID NO: 6.

5 En otra realización preferente, la célula de la presente invención comprende un vector de expresión que comprende a su vez una secuencia nucleotídica según SEQ ID NO: 8.

En un quinto aspecto, la presente invención se refiere a un sistema CRISPR-Cas que comprende un ARN guía y la proteína de la presente invención (de aquí en adelante, “sistema CRISPR-Cas de la presente invención”).

10

En el ámbito de interpretación de la presente invención, se entenderá por “sistema CRISPR-Cas” cualquier sistema que comprenda los elementos implicados en la expresión y/o actividad de los genes asociados a dicho sistema, incluyendo tanto la/s secuencia/s nucleotídica/s que se transcribe/n para generar el ARN guía como la secuencia

15

nucleotídica que codifica la proteína de la presente invención.

Asimismo, se entenderá por “ARN guía”, cualquier construcción de ARN monocatenaria, bicatenaria o parcialmente bicatenaria que se asocie con la proteína de la presente invención y que comprenda una secuencia ribonucleotídica complementaria a una

20

secuencia de ADN específica de una célula (“secuencia diana”); esto es, que forma enlaces de hidrógeno con las bases nitrogenadas de los nucleótidos de la secuencia diana.

Preferentemente, dicho ARN guía comprende dos moléculas de ARN, tracrRNA y crRNA, parcialmente complementarias entre sí formando el híbrido tracrRNA:crRNA, o una

25

molécula de ARN parcialmente bicatenaria (sgRNA).

En una realización preferente, el sistema CRISPR-Cas de la presente invención comprende una proteína con una secuencia aminoacídica con al menos un 70% de identidad de secuencia con SEQ ID NO: 1.

30

En otra realización preferente, el sistema CRISPR-Cas de la presente invención comprende una proteína con una secuencia aminoacídica según SEQ ID NO: 5.

En otra realización preferente, el sistema CRISPR-Cas de la presente invención

35

comprende una proteína con una secuencia aminoacídica según SEQ ID NO: 7.

En otra realización preferente, el sistema CRISPR-Cas de la presente invención comprende un ARN guía procedente de la transcripción de una secuencia nucleotídica según SEQ ID NO: 3.

5 En un sexto aspecto, la presente invención se refiere al uso de la proteína de la presente invención, y/o de la secuencia nucleotídica de la presente invención, y/o del vector de expresión de la presente invención, y/o de la célula de la presente invención, y/o del sistema CRISPR-Cas de la presente invención para:

- la modificación genética, regulación de expresión génica y/o visualización *in vivo*
- 10 de secuencias nucleotídicas concretas; y/o
- el diagnóstico molecular de enfermedades; y/o
- la producción de antimicrobianos específicos de secuencia.

Preferentemente, para la modificación genética, regulación de expresión génica y/o

15 visualización *in vivo* de secuencias nucleotídicas concretas de células eucariotas.

Alternativamente, para la producción de antibacterianos. Preferentemente, para la producción de antibacterianos frente a *Escherichia coli*.

20 DESCRIPCIÓN DE LAS FIGURAS

La Figura 1 muestra una representación esquemática del locus CRISPR-EHCas9 y los dominios de la proteína EHCas9 (de aquí en adelante, proteína de la presente invención). El locus CRISPR-EHCas9 comprende tres genes *cas* en el orden *cas9* (denominado

25 *ehcas9*) - *cas1* - *cas2* (representados mediante rectángulos apuntando en el sentido de la transcripción) y dos unidades CRISPR de 36 pares de bases (pb; rectángulos blancos) separados por un espaciador de 29 pb (rombo). La ubicación de un posible gen del *tracrRNA* se representa como una flecha que apunta en el sentido de la transcripción. El gen *ehcas9* codifica la proteína de la presente invención, cuya estructura comprende los

30 siguientes dominios: RuvC (motivos I, II y III), Bridge Helix (BH), de reconocimiento (REC), nucleasa HNH, Phosphate Lock Loop (PLL), WED y de interacción con PAM (PI).

La Figura 2A muestra el alineamiento de la secuencia de la proteína de la presente invención, SEQ ID NO: 1, con la del ortólogo estructuralmente caracterizado más cercano,

35 correspondiente a *Corynebacterium diphtheriae* (CdCas9; ID de la base de datos de proteínas 6JOO). Los límites de los dominios RuvC (motivos RuvCI-III), Bridge Helix (BH),

reconocimiento (REC), HNH, Phosphate Lock Loop (PLL), WED y de interacción con PAM (PI) de CdCas9 se indican mediante barras debajo de la secuencia. La Figura 2B muestra el alineamiento múltiple de SEQ ID NO: 1 con la secuencia de los ortólogos estructuralmente caracterizados: CjCas9, *Campylobacter jejuni*; NmCas9, *Neisseria meningitidis* 8013; StCas9, *Streptococcus thermophilus* LMD9; SaCas9, *Staphylococcus aureus*; SpCas9, *Streptococcus pyogenes*. Se enumeran algunas de las posiciones de aminoácidos de SEQ ID NO: 1. El sitio catalítico RuvC está sombreado y el sitio catalítico HNH se muestra en negrita y subrayado. En ambas figuras, las posiciones conservadas están marcadas con un asterisco.

10

La Figura 3 muestra la relación evolutiva de la proteína de la presente invención mediante el árbol filogenético de dicha proteína y 798 proteínas ortólogas. Los clados II, III, IV y V pertenecen al subtipo II-A, el clado I al subtipo II-B y los clados VI, VII, VIII, IX y X al subtipo II-C. Cas9 de *Sulfitobacter donghicola* (SdoCas9) y ortólogos comúnmente utilizados para la edición del genoma (SaCas9: *Staphylococcus aureus* Cas9; SpCas9: *Streptococcus pyogenes* Cas9; NmCas9: *Neisseria meningitidis* Cas9; CjCas9: *Campylobacter jejuni* Cas9; CdCas9: *Corynebacterium diphtheriae* Cas9; StCas9: *Streptococcus thermophilus* Cas9) están etiquetados en su posición aproximada en el árbol.

15

20

La Figura 4 muestra los mapas de los principales plásmidos construidos en esta invención. La Figura 4A muestra el mapa del plásmido pMML02, que incluye un gen que confiere resistencia a cloranfenicol (CmR), un gen que codifica la proteína de la presente invención bajo el control del promotor pBAD, y una agrupación CRISPR constituida por dos repeticiones de 36 pb y un espaciador de 29 pb con diana en el plásmido pSEVA. La transcripción de la agrupación CRISPR está controlada por un promotor constitutivo (Part:BBa_J23101) y finaliza en un terminador artificial (Part:BBa_B1006). La Figura 4B muestra el mapa del plásmido pMML03, que incluye un gen que confiere resistencia a ampicilina (AmpR) y la región intergénica del sistema CRISPR-Cas9 de la presente invención donde se localiza la secuencia codificante de EH tracrRNA bajo el control del promotor de lactosa Part:BBa_R0010 (promotor lac). La Figura 4C muestra el mapa del plásmido pMML09, que incluye un gen que confiere resistencia a cloranfenicol (CmR), el gen que codifica la proteína de la presente invención bajo el promotor PBAD (Part:BBa_I0500), y la región codificante de EH sgRNA. La región espaciadora del EH sgRNA coincide con una secuencia del gen cromosómico *pyrF* de *E. coli*. La transcripción del gen de EH sgRNA está controlada por un promotor constitutivo (Part:BBa_J23101) y finaliza en un terminador artificial (Part:BBa_B1006). La Figura 4D muestra el mapa del

25

30

35

plásmido pMML12, que incluye genes que confieren resistencia a kanamicina (KanR) y a ampicilina (AmpR) y el gen que codifica la proteína de la presente invención con uso de codones optimizado para células humanas (EHCas9 humanizada), fusionado a una secuencia codificante de una señal de localización nuclear (SV40 NLS), bajo el control del promotor del citomegalovirus humano (promotor CMV, Part:BBa_K2605001). La Figura 4E muestra el mapa del plásmido pMML13, que incluye un gen que confiere resistencia a ampicilina (AmpR) y un gen bajo el control de un promotor U6 (R1LP2N), que codifica la región constante de EH sgRNA y una zona espaciadora intercambiable, flanqueada por sitios de restricción Esp3I. La Figura 4F muestra el mapa del plásmido pMML22, que incluye un gen que confiere resistencia a kanamicina (KanR) y un gen con uso de codones optimizado para *E. coli* que codifica la proteína de la presente invención fusionada en el extremo N-terminal con una cola de 6 histidinas (6xHis). La transcripción del gen *ehcas9-6xhis* está bajo el control de un promotor T7 (Part:BBa_I719005).

La Figura 5 muestra el cribado y validación de PAM. (A) Logotipo de secuencia de la región PAM preferida por la proteína de presente invención para la escisión de la diana, según lo determinado por el cribado *in vivo* de una genoteca PAM. Se indican las posiciones de los nucleótidos desde el extremo 3' de la secuencia diana (cadena coincidente con el espaciador). Se ensayaron los nucleótidos de las posiciones 2 a 4 (la primera posición se mantuvo invariable, correspondiendo a timina). (B) Logotipo de la secuencia PAM consenso preferido por la proteína de la presente invención para la escisión de la diana según lo determinado mediante cribado *in vitro*. Se indican las posiciones de los nucleótidos desde el extremo 3' de la secuencia diana. En este caso se ensayaron los nucleótidos desde la posición 1 a la 7. (C) Validación de PAM *in vivo*. La eficiencia de transformación (número de unidades formadoras de colonias - CFU - por µg de ADN plasmídico) de células de *E. coli* que expresan (+ EHCas9) o no (- EHCas9) la proteína de la presente invención además de un EH crRNA guía y el EH tracrRNA predicho, con plásmidos que portan una diana adyacente a secuencias que varían en las posiciones 2, 3 y 4 (ACC, GGA, GGC, GGG, GGT) de la región PAM. Los datos son la media de tres réplicas (las barras de error corresponden a la desviación estándar).

La Figura 6 muestra el esquema del EH sgRNA que incluye un espaciador genérico de 23 nucleótidos (nt) apareado con la hebra diana en un sustrato de ADN que contiene una secuencia coincidente con el espaciador y un PAM compatible (en cursiva). La secuencia de EH tracrRNA, que comprende el conector (tetraloop 5'-GAAA-3', subrayado), la anti-

repetición y los dos segmentos formadores de tallo-bucle está resaltada en negrita, y la secuencia de la región de la repetición se encuentra recuadrada.

La Figura 7 muestra el resultado de la electroforesis en gel de poliacrilamida SDS de los pasos de la purificación de la proteína de la presente invención fusionada a una cola de 6 histidinas (EHCas9-6xHis; SEQ ID NO: 5). Se incluye un lisado de bacterias que expresan EHCas9-6xHis (Lisado) y muestras de extractos de proteínas purificadas a través de la columna de unión a His (Columna His), así como después de la filtración en gel (Filtración en gel). Se indica el tamaño de las bandas correspondientes a un marcador de peso molecular de proteínas (M). La banda principal de los extractos de proteínas corresponde a una proteína de alrededor de 120 kDa.

La Figura 8 muestra los resultados de la electroforesis en gel de agarosa de productos de reacción de la proteína de la invención obtenidos mediante ensayos de digestión *in vitro* con sustratos de ADN bicatenarios. Por defecto, las reacciones se llevaron a cabo bajo las siguientes condiciones estándar: durante 30 min a +37°C en presencia de 20 mM MgCl₂ y 25 nM de ADN diana con PAM 5'-NGG-3', tras añadir una solución con EHCas9 (0,5 µM) y EH sgRNA (0,5 µM) previamente incubada (Preincubado) durante 15 min a +37°C. Se indica el tamaño de bandas relevantes de un marcador de peso molecular de ADN (M, en kpb) y la posición correspondiente al sustrato de ADN sin cortar, así como las de los dos fragmentos que se generarían tras su digestión (cortado). (A) Muestras de reacciones de digestión en condiciones estándar utilizando todos los componentes de la reacción con el complejo EHCas9:EH sgRNA preincubado (carril 2) o sin preincubar (carril 7), y en ausencia de algún componente (MgCl₂, carril 3; diana con PAM, carril 4; EH sgRNA, carril 5; EHCas9, carril 6), tras preincubar (Preincubado; carriles 3 y 4) o sin preincubación (carriles 5 y 6). (B) Muestras de reacciones de digestión en condiciones estándar con diferentes concentraciones de proteína. (C) Muestras de las reacciones de digestión en condiciones estándar salvo por el tiempo de incubación. (D) Muestras de las reacciones de digestión en condiciones estándar excepto para la temperatura de incubación.

La Figura 9 se refiere a la edición genética de *E. coli* asistida por la proteína de la presente invención. La Figura 9A muestra el esquema del procedimiento para la selección positiva de mutantes de *E. coli* obtenidos tras recombinación genética. La Figura 9B muestra el resultado de la electroforesis en gel de agarosa de productos de PCR obtenidos a partir de colonias de transformantes obtenidas en experimentos de edición del gen *pyrF* (GDI). Las colonias provienen de la co-transformación de un molde de recombinación (la

recombinación daría lugar a una delección de 0,6 kpb en *pyrF*), y un plásmido que codifica para EHCas9 y un EH sgRNA dirigido a una secuencia diana en el gen *pyrF* (+EHCas9) o con un plásmido equivalente pero que solo codifica el EH sgRNA (-EHCas9). Cada carril se corresponde con una colonia elegida al azar. Se señala el tamaño de bandas relevantes de un marcador de peso molecular de ADN (M, en kpb) y las posiciones esperadas para las bandas correspondientes al amplicón del gen *pyrF* original (ca. 1 kpb; Silvestre) y el del gen con la delección (ca. 0,5 kpb; Mutante).

La Figura 10 se refiere a la edición genética en cultivos de células N2a de ratón asistida por la proteína de la presente invención. La Figura 10A muestra el esquema del procedimiento de edición genética. La Figura 10B muestra el porcentaje del número de lecturas de secuenciación con inserciones o delecciones (% INDELS; n=3, media $\pm$ s.d.) obtenidos para 4 secuencias diana en los genes *Lrmda* (*Lrmda*.1) y *Oca2* (*Oca2*.2, *Oca2*.3, *Oca2*.4) del genoma de ratón tras la transfección con plásmidos que codifican para SpCas9 y Sp sgRNA (SpCas9.sgRNA; a), EHCas9 y EH sgRNA (EHCas9.sgRNA; b) o EHCas9 (EHCas9; c). Como control negativo se incluyen los resultados obtenidos con células no transfectadas (N2a; d). La Figura 10C muestra el alineamiento de los 10 alelos más frecuentes revelados para la diana *Oca2*.3 en los experimentos con sistema CRISPR-EHCas9 de la invención. Los códigos de las delecciones figuran en la columna de la izquierda (p. ej., -2:1D, delección de un nucleótido en la posición -2 respecto al sitio de corte). La posición del sitio de corte preferente de EHCas9 se muestra con una línea discontinua. La secuencia de la región diana original (*Oca2*.3) se incluye en la primera línea. Las regiones diana y PAM se marcan con letras subrayadas y en un recuadro, respectivamente. En la columna de la derecha se representa la frecuencia de cada alelo (%) como porcentaje promedio de 3 réplicas.

La Figura 11 muestra el crecimiento de células N2a de ratón expresando componentes de la proteína de la presente invención, EHCas9, y de la proteína Cas9 de la bacteria *Streptococcus pyogenes*, SpCas9. (A) Recuento de células nucleadas sin transfectar (Sin plásmido) y transfectadas con 200 ng, 150 ng o 100 ng de plásmidos codificando SpCas9 o EHCas9. (B) Recuento de núcleos en células transfectadas y sin transfectar (Sin plásmido) con 100 ng de plásmidos codificando sgRNA de SpCas9 (Sp sgRNA) o EHCas9 (EH sgRNA) (n=3, media $\pm$ DS). (C) Tinción DAPI de células no transfectadas (N2a) y de células transfectadas con plásmidos codificando SpCas9 o EHCas9.

DESCRIPCIÓN DETALLADA DE LA INVENCION

La presente invención proporciona una proteína endonucleasa Cas9 con un tamaño tal que facilita su administración tanto a bacterias como células de mamíferos mediante vectores comúnmente empleados en biotecnología y biomedicina para la edición génica de las mismas. Ventajosamente, la proteína de la presente invención comprende una secuencia aminoacídica según SEQ ID NO:1 (de aquí en adelante, "EHCas9"). Preferentemente, una secuencia aminoacídica con al menos un 70% de identidad de secuencia con SEQ ID NO: 1. Aún más preferentemente, una secuencia aminoacídica según SEQ ID NO: 5 o SEQ ID NO: 7.

Las cepas bacterianas, los plásmidos, y los oligonucleótidos utilizados en los ejemplos de la presente invención son los que se relacionan en las Tablas 1, 2 y 3, respectivamente.

Tabla 1. Cepas de *E. coli* utilizadas en la presente invención.

Cepa	Genotipo relevante	Uso	Referencia
NZYstar	<i>endA1 hsdR17(rk-, mk+) supE44 thi -1 recA1 gyrA96 relA1 lac[F' proA+B+ lacIq ZDM15:Tn10(TcR)]</i>	Clonación de plásmidos y generación de biblioteca PAM	NZYTech
BW 27783	<i>lacIq rmB3 ΔlacZ4787 Δ(araBAD)567 Δ(rhaBAD)568 hsdR514 Δ(K) Φ(ΔaraEp PCP8-araE)</i>	Cribado <i>in vivo</i> y validación de PAM. Selección positiva de células de <i>E. coli</i> editadas genéticamente.	[1]
TOP10	<i>F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 nupG recA1 araD139 Δ(ara-leu)7697 galE15 galK16 rpsL(StrR) endA1 λ-</i>	Clonación de plásmidos utilizados en edición genética en eucariotas	Invitrogen
BL21(DE3)	<i>F- ompT gal dcm lon HsdSB (rB-mB-) λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])</i>	Producción de EHCas9	[2]

[1] Khlebnikov, A. et al (2001). Microbiology, 147, 3241-3247; [2] Rosenberg, A.H. et al. (1987). Gene, 56, 125-135.

Tabla 2. Plásmidos utilizados en la presente invención.

Plásmido	Descripción/Uso	Referencia
pBAD33	Vector de expresión en bacterias bajo el promotor de arabinosa. Resistencia a cloranfenicol	[3]
pUC57	Vector de clonación bacteriana. Resistencia a ampicilina	[4]
pSEVA431	Vector de clonación bacteriana. Resistencia a espectinomicina	[5]
pHTP1	Vector de expresión en bacterias bajo el promotor T7 con una cola 6x His. Resistencia a kanamicina	NZYTEch
pKD46	Plásmido bacteriano de replicación termosensible codificante del sistema de recombinación Lambda Red. Resistencia a ampicilina	[6]
hCas9	Expresión de SpCas9 humanizado para edición genómica. Resistencia a ampicilina y a la kanamicina	[7]
MLM3636	Vector de expresión de SpCas9 sgRNA expression. Resistencia a ampicilina	Keith Joung (no publicado)
pUC57-EHCas9	pUC57 que codifica EHCas9	Esta invención
pUC57-EHCRISPR	pUC57 que codifica la agrupación EHCas9 CRISPR	Esta invención
pMML01	Derivado de pBAD33 que contiene la agrupación EH CRISPR	Esta invención
pMML02	Derivado de pMML01 que contiene el gen <i>ehcas9</i>	Esta invención
pMML03	Derivado de pUC57 que contiene regiones intergénicas de CRISPR-EHCas9	Esta invención
pMML04	Derivado de pSEVA431 que contiene una diana y la secuencia PAM 5'-TGGA-3'	Esta invención
pMML05	Derivado de pSEVA431 que contiene una diana y la secuencia PAM 5'-TGGC-3'	Esta invención
pMML06	Derivado de pSEVA431 que contiene una diana y la secuencia PAM 5'-TGGG-3'	Esta invención
pMML07	Derivado de pSEVA431 que contiene una diana y la secuencia PAM 5'-TG GT-3'	Esta invención
pMML08	Derivado de pUC57 que contiene la secuencia que codifica EH sgRNA sin el espaciador	Esta invención

Plásmido	Descripción/Uso	Referencia
pMML09	Derivado de pMML02 en el que se sustituye la agrupación EH CRISPR por una secuencia que codifica un EH sgRNA que contiene un espaciador que coincide con una secuencia en <i>pyrF</i>	Esta invención
pMML10	Derivado de pMML09 por delección de <i>ehcas9</i>	Esta invención
pMML11	Derivado de pUC57 que contiene el gen <i>ehcas9</i> con uso de codones optimizado para humanos	Esta invención
pMML12	Derivado de hCas9 que contiene el gen <i>ehcas9</i> con uso de codones optimizado para humanos de pMML11 en lugar de <i>hcas9</i>	Esta invención
pMML13	Derivado de MLM3636 que codifica EH sgRNA (sin espaciador) de pMML08	Esta invención
pMML14	Derivado de MLM3636 por inserción de un espaciador dirigido al locus <i>Oca2.2</i>	Esta invención
pMML15	Derivado de MLM3636 por inserción de un espaciador dirigido al locus <i>Oca2.3</i>	Esta invención
pMML16	Derivado de MLM3636 por inserción de un espaciador dirigido al locus <i>Oca2.4</i>	Esta invención
pMML17	Derivado de MLM3636 por inserción de un espaciador dirigido al locus <i>Lrmda.1</i>	Esta invención
pMML18	Derivado de pMML13 por inserción de un espaciador dirigido al locus <i>Oca2.2</i>	Esta invención
pMML19	Derivado de pMML13 por inserción de un espaciador dirigido al locus <i>Oca2.3</i>	Esta invención
pMML20	Derivado de pMML13 por inserción de un espaciador dirigido al locus <i>Oca2.4</i>	Esta invención
pMML21	Derivado de pMML13 por inserción de un espaciador dirigido al locus <i>Lrmda.1</i>	Esta invención
pMML22	Derivado de pHTP1 que contiene el gen <i>ehcas9</i> con uso de codones optimizado para <i>E. coli</i>	Esta invención

[3] Guzman, L.-M. et al. (1995). J. Bacteriol., 177, 4121–4130; [4] Yanisch-Perron, C. et al. (1985). Gene, 33, 103–119; [5] Silva-Rocha, R. et al. (2013). Nucleic Acids Res., 41, D666–D675; [6] Datsenko, K.A. and Wanner, B.L. (2000). Proc Natl Acad Sci U S A, 97, 6640–6645; [7] Mali, P. et al. (2013). Science, 339, 823–826.

Tabla 3. Oligonucleótidos empleados en la presente invención.

#ID	Objetivo	Diana	Secuencia (5'→3')	SEQ ID NO
8	Construcción de pMML01	pUC57-EHCRISPR	GAAGCATTATCAGGGTTATCGCACTCGAGT TTACAGCTAGC	SEQ ID NO: 9
9	Construcción de pMML01	pUC57-EHCRISPR	ATGTATCCGCTCATGAGACATGCGCTCGAGA AAAAAACCCCG	SEQ ID NO: 10
10	Construcción de pMML01	pBAD33	TAGCTGTAAACTCGAGTGCATAACCCCTGAT AAATGCTTCAATAATATTGAAAAAGGAAG	SEQ ID NO: 11
11	Construcción de pMML01	pBAD33	GGTTTTTTTCTCGAGCGCATGTCTCATGAG CGGATACATAATTTGAATGT	SEQ ID NO: 12
20	Construcción de pMML02	pUC57-EHCas9	TTTAAGAAAGGAGATATACATATGACCCATAAT GCATATACCTTGGG	SEQ ID NO: 13
21	Construcción de pMML02	pUC57-EHCas9	GCCTGCAGGTCGACTCTAGATTACGGGCGC GTACGC	SEQ ID NO: 14
22	Construcción de pMML02	pMML01	GTCCGCGTACGCGCCCGTAATCTAGAGTCG ACCTGCAGGCATGCAAGCT	SEQ ID NO: 15
23	Construcción de pMML02	pMML01	GTATATGCATTATGGTGCATATGTATATCTCC TTCTTAAAGGTACCGAGCT	SEQ ID NO: 16
107	Generación de una genoteca PAM derivada de pSEVA31 para cribado in vivo de PAM	pSEVA431	TCATTATAGCGATTTTTCNNNATATCCATCC TTTTTCGCACG	SEQ ID NO: 17
108	Generación de una genoteca PAM derivada de pSEVA31 para cribado in vivo de PAM	pSEVA431	CGTGCGAAAAAGGATGGATATNNGAAAAAA TCGCTATAATGA	SEQ ID NO: 18
159	Construcción de pMML04-07	pSEVA431	CCCCGAAGCAGGGTTATG	SEQ ID NO: 19
134	Construcción de pMML05	pSEVA431	TCATTATAGCGATTTTCTCCATATCCATCC TTTTTCGCACG	SEQ ID NO: 20
133	Construcción de pMML05	pSEVA431	TCATTATAGCGATTTTTCGCCATATCCATCC TTTTTCGCACG	SEQ ID NO: 21

#ID	Objetivo	Diana	Secuencia (5'→3')	SEQ ID NO
187	Construcción de pMML06	pSEVA431	TCATTATAGCGATTTTTCCTCCCATATCCATCC TTTTTCGCACG	SEQ ID NO: 22
186	Construcción de pMML07	pSEVA431	TCATTATAGCGATTTTTCACCCATATCCATCC TTTTTCGCACG	SEQ ID NO: 23
70	Amplificación de la región PAM para su secuenciación (cribado PAM)	pSEVA431	GCGAATAAGGGACAGTGAA	SEQ ID NO: 24
60	Amplificación de la región PAM para su secuenciación (cribado PAM)	pSEVA431	GGCCCAGTATCAGCCCCGTC	SEQ ID NO: 25
169	Generación del molde de ADNbc para transcripción in vitro de EH sgRNA (ensayos de digestión)	pMML08	AAGCTAATACGACTCACTATAGGGTATCGTG CGAAAAAGGATGGATAGCTGGGCTGGACC ACGA	SEQ ID NO: 26
170	Generación del molde de ADNbc para transcripción in vitro de EH sgRNA (ensayos de digestión)	pMML08	ACAAAAGAGGGCAAGGC	SEQ ID NO: 27
115	Generación del ADNbc diana para los ensayos de digestión	pSEVA431	ATTCTTGCAGGTATCTTCGAGCC	SEQ ID NO: 28
15	Generación del ADNbc diana para los ensayos de digestión	pSEVA431	CCAGCCTTGACCGAAACG	SEQ ID NO: 29
218	Construcción de pMML09	pMML08	CTGCTGCAGCTGACTTAGGCGTGGCTGCGG CTGGACCCACGA	SEQ ID NO: 30
170	Construcción de pMML09	pMML08	ACAAAAGAGGGCAAGGC	SEQ ID NO: 27

#ID	Objetivo	Diana	Secuencia (5'→ 3')	SEQ ID NO
203	Generación del intermediario de ADNbc para la construcción de pMML09	pMML02	TTTTTTTTTATTATTATTAAACAAAAGAGGGC AAGGCCG	SEQ ID NO: 31
219	Generación del intermediario de ADNbc para la construcción de pMML09	pMML02	ATTATGCTAGCATCCTAGATCTGCTGCAGCT GACTTAGG	SEQ ID NO: 32
224	Construcción de pMML09	Intermediario de ADNbc para la construcción de pMML09	CACGCCTAAGTCAGCTGCAGCAGATCTAGG ATGCTAGCATAATACCTAGGACT	SEQ ID NO: 33
202	Construcción de pMML09	Intermediario de ADNbc para la construcción de pMML09	CCGGCCTTGCCCTCTTTTGTTAATAATAATAA AAAAAAAACCCGCCCTGACAGGGC	SEQ ID NO: 34
22	Construcción de pMML10	pMML09	GTCGCGGTACGCGCCCGTAATCTAGAGTCG ACCTGCAGGCATGCAAGCT	SEQ ID NO: 35
23	Construcción de pMML10	pMML09	GTATATGCATTATGGGTCATATGTATATCTCC TTCCTAAAGGTACCGAGCT	SEQ ID NO: 36
188	Generación del molde de recombinación de ADNbc	<i>E. coli</i> BW 27783	TCCATTGTGCTCCGGCAAC	SEQ ID NO: 37
189	Generación del molde de recombinación de ADNbc	<i>E. coli</i> BW 27783	CTGTTCTGGCGTCATAATGCGACCAGACCTT CTTGATGATGG	SEQ ID NO: 38
190	Generación del molde de recombinación de ADNbc	<i>E. coli</i> BW 27783	ATCATCAAGAAGGTCTGGTCGCATTATGACG CCAGAAACA	SEQ ID NO: 39

#ID	Objetivo	Diana	Secuencia (5'→ 3')	SEQ ID NO
191	Generación del molde de recombinación de ADNbc	<i>E. coli</i> BW 27783	TGAGTAGACCAGACGGCTG	SEQ ID NO: 40
234	Cribado de selección positiva del gen pyrF	Colonias derivadas de células editadas de <i>E. coli</i>	TACTGGCATTGTCCGCTTGTGTC	SEQ ID NO: 41
235	Cribado de selección positiva del gen pyrF	Colonias derivadas de células editadas de <i>E. coli</i>	GCACCACACCGTCGCCT	SEQ ID NO: 42
P1	Construcción de pMML12	pMML11	TAGAGGATCGAACCCCTTGCCACCATGGCCC ACAATGCTTATACGC	SEQ ID NO: 43
P2	Construcción de pMML12	pMML11	TTCTTCTTGGGGTCAGCCCTGCTTGGGCGC GTTCTGGGAC	SEQ ID NO: 44
P3	Construcción de pMML12	hCas9	GATCCGGGTCCCAGAACGCCCAAGCAGG GCTGACCCCAAG	SEQ ID NO: 45
P4	Construcción de pMML12	hCas9	GCGTATAAGCATTGTGGCCCATGGTGCAA GGGTTCGATCCTC	SEQ ID NO: 46
P5	Construcción de pMML13	pMML08	ACCGAGACGATTAATGCGTCTCGCTGCGGC TGGACCCACGAG	SEQ ID NO: 47
P6	Construcción de pMML13	pMML08	CGAGCGGCCCAAGCTTAAAAAACAACAAAGA GGCAAGGCCGAG	SEQ ID NO: 48
P7	Construcción de pMML13	MLM3636	GCTCCGGCCTTGCCCTCTTTTGTTTTTTTAA GCTTGGGCCGCTCG	SEQ ID NO: 49
P8	Construcción de pMML13	MLM3636	GATTCTCGTGGTCCAGCCGCGCAGCGAGACG CATTAAATCGTCTCGG	SEQ ID NO: 50
P17	Construcción de pMML15 y pMML19 por Golden Gate	P18 (pMML15) y	ACACCTACCAAGGCTGCTGTGTGGAGAG	SEQ ID NO: 51

#ID	Objetivo	Diana	Secuencia (5'→ 3')	SEQ ID NO
		P19 (pMML19)		
P18	Construcción de pMML15 por Golden Gate	P17	GCAGCTCTCCACACACAGCAGCCCTTTGGTAG	SEQ ID NO: 52
P19	Construcción de pMML19 por Golden Gate	P17	AAACTCTCCACACACAGCAGCCTTTGGTAG	SEQ ID NO: 53
P20	Construcción de pMML14 y pMML18 por Golden Gate	P20 (pMML14) y P21 (pMML18)	ACACCGCTTGTCAAGTGCCTGACGGTGCG	SEQ ID NO: 54
P21	Construcción de pMML14 por Golden Gate	P20	GCAGCGCACCCGTCAGGCACCTTGACAAGCG	SEQ ID NO: 55
P22	Construcción de pMML18 por Golden Gate	P20	AAACGCGACCCGTCAGGCACCTTGACAAGCG	SEQ ID NO: 56
P23	Construcción de pMML17 y pMML21 por Golden Gate	P24 (pMML17) y P25 (pMML21)	ACACCCAGAACTGGGAATAGTTTGGGCTG	SEQ ID NO: 57
P24	Construcción de pMML17 por Golden Gate	P23	GCAGCAGCCCCAAACTATTCCCAGTTCTGG	SEQ ID NO: 58
P25	Construcción de pMML21 por Golden Gate	P23	AAACAGCCCCAAACTATTCCCAGTTCTGG	SEQ ID NO: 59
P26	Construcción de pMML16 y pMML20 por Golden Gate	P27 (pMML16) y P28 (pMML20)	ACACCATTTGGTGGGTCCCCCAATAGCAGG	SEQ ID NO: 60

#ID	Objetivo	Diana	Secuencia (5'→ 3')	SEQ ID NO
P27	Construcción de pMML16 por Golden Gate	P26	GCAGCCTGCTATTGGGGACCCACCAAATG	SEQ ID NO: 61
P28	Construcción de pMML20 por Golden Gate	P26	AAACCTGCTATTGGGGACCCACCAAATG	SEQ ID NO: 62
P45	Amplificación de la región diana Oca2.2 en ensayos de edición	Células N2a sometidas a ensayos de edición	GGAAACACTTATGGTGATTCGAG	SEQ ID NO: 63
P46	Amplificación de la región diana Oca2.3 en ensayos de edición	Células N2a sometidas a ensayos de edición	GCACAGGATTCAGACAGG	SEQ ID NO: 64
P47	Amplificación de la región diana <i>Lrmda.1</i> en ensayos de edición	Células N2a sometidas a ensayos de edición	CGTGAAGGATGCGGTGAGTG	SEQ ID NO: 65
P48	Amplificación de la región diana Oca2.4 en ensayos de edición	Células N2a sometidas a ensayos de edición	ATTTGGGTTGTGACGGACC	SEQ ID NO: 66
P49	Amplificación de la región diana Oca2.2 en ensayos de edición	Células N2a sometidas a ensayos de edición	GCAGCCTTTGGTACAGTTATC	SEQ ID NO: 67
P50	Amplificación de la región diana Oca2.3 en ensayos de edición	Células N2a sometidas a ensayos de edición	CCCTTCTCACATTGAGCCCT	SEQ ID NO: 68

#ID	Objetivo	Diana	Secuencia (5'→ 3')	SEQ ID NO
P51	Amplificación de la región diana <i>Lrmda.1</i> en ensayos de edición	Células N2a sometidas a ensayos de edición	GTCGTCTCCGAGCAGATTG	SEQ ID NO: 69
P52	Amplificación de la región diana <i>Oca2.4</i> en ensayos de edición	Células N2a sometidas a ensayos de edición	CCATCACGTTACTGTCAATACAGG	SEQ ID NO: 70

A menos que se especifique lo contrario, los cultivos de *E. coli* se crecieron a +37°C en medio líquido *Luria-Bertani* (LB) con agitación orbital a 180 rpm, o en LB agar. Para la selección de células portadoras de plásmidos, los medios se suplementaron con cloranfenicol (25 µg/ml), ampicilina (100 µg/ml), espectinomicina (50 µg/ml) o kanamicina (50 µg/ml), según correspondiera.

Las secuencias espaciadoras guía se clonaron en el plásmido pMML13 (Figura 4E) usando el método Golden Gate [Engler, C. et al. (2009). PLOS ONE, 4, e5553]. Los otros ensayos de clonación molecular y sustitución de genes de plásmidos fueron realizados mediante ensamblaje Gibson con el kit de clonación Gibson Assembly® (NEB).

Para la preparación de células electrocompetentes de *E. coli* BL21(DE3) y *E. coli* BW27783, cultivos líquidos en fase estacionaria se llevaron a una dilución 1/100 en caldo LB y se crecieron hasta una $DO_{600} = 0,5$. Las células fueron recogidas por centrifugación y lavadas tres veces con agua desionizada y una vez con glicerol al 10%. Las transformaciones se realizaron con 50 µl de suspensiones de células electrocompetentes recién preparadas, incubadas en hielo durante 25 min después de agregar el ADN. La mezcla de células y ADN se transfirió a una cubeta de electroporación de tamaño de ranura de 2 mm (Molecular Bioproducts) enfriada sobre hielo y se sometió a electroporación a 2,5 kV con un MicroPulser (BIORAD). Inmediatamente a continuación se añadió a la suspensión celular 1 ml de caldo SOC y se incubó durante 1 hora en condiciones estándar en un tubo de 12 ml. Finalmente, las células se sembraron en placas con medios suplementados con el antibiótico correspondiente para la selección del plásmido y se incubaron durante toda la noche a +30°C en el caso del plásmido termosensible pKD46 o a +37°C en el resto de los casos.

Las células de *E. coli* NZYStar (NZYTech) y *E. coli* TOP10 (Invitrogen) químicamente competentes se transformaron siguiendo las instrucciones del fabricante.

Los plásmidos se aislaron de *E. coli* con el kit PureLink™ HiPure Plasmid Midiprep o el kit PureLink™ HiPure Plasmid Miniprep (Invitrogen). Los productos de PCR y los fragmentos de ADN se purificaron con el GFX™ PCR DNA and Gel Band Purification Kit (Cytiva).

La concentración y la pureza de las soluciones de ácidos nucleicos se estimaron con un espectrofotómetro NanoDrop ND-1000 (Thermo Scientific), y su integridad se evaluó mediante electroforesis en gel de agarosa.

Para visualizar las moléculas de ADN sometidas a electroforesis en gel de agarosa, se tomaron imágenes de geles que contenían GreenSafe premium (NZYTech), empleando ChemiDoc XRS+ Gel Imaging System (BIORAD). El marcador de peso molecular 1 Kb Plus DNA Ladder (Invitrogen) se incluyó en los geles de agarosa para estimar el tamaño de los fragmentos de ADN.

Ejemplo 1: Identificación y caracterización de la proteína de la invención con secuencia aminoacídica según SEQ ID NO:1

Para la identificación y caracterización de la proteína de la presente invención con secuencia aminoacídica según SEQ ID NO: 1 (de aquí en adelante, 'EHCas9'), los inventores recolectaron muestras de agua en una laguna del Parque Natural 'El Hondo' (EH) en España. Estas muestras se prefiltraron a través de papel de filtro y un filtro de membrana Durapore® de 5 µm de tamaño de poro (Merk). Posteriormente, se realizó una filtración secuencial a través de un filtro de membrana Durapore® (Merk) de 0,22 µm de tamaño de poro y un dispositivo de ultrafiltración de flujo cruzado VIVAFLOW 200 de 30.000 MWCO (Sartorius). La muestra filtrada se concentró utilizando un filtro 3K Ultra Amicon® (Millipore). El ADN se purificó a partir del concentrado con el kit PureLink® Viral RNA/DNA Mini (Invitrogen).

La secuenciación del ADN fue realizada utilizando Illumina HiSeq. Las lecturas de baja calidad fueron eliminadas con el programa PRINSEQ-lite [Schmieder, R., & Edwards, R. (2011). *Bioinformatics*, 27(6), 863-864], utilizando la configuración: min_length: 50, trim_qual_right: 30, trim_qual_type: mean y trim_qual_window: 20. Seguidamente, las secuencias eucariotas se identificaron mediante búsquedas BLASTn (opciones: -taxidlist: taxid:2759, -evalue: 0.005) frente a la base de datos del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés; <https://blast.ncbi.nlm.nih.gov/Blast.cgi>). Las secuencias con una identidad superior a 0,9 se filtraron usando el script FastQ.filter.pl de Enveomics Collection [Rodriguez-R, L.M. & Konstantinidis, K.T. (2016). *PeerJ Preprints*, 4, e1900v1]. El ensamblaje *de novo* de las lecturas restantes se realizó con el programa SPAdes v3.13.0 [Nurk, S., et al. (2017). *Genome Res.*, 27(5), 824-834] usando la opción metaspades con parámetros: -k 21, 33, 55, 77, 99, 127.

Para la identificación de los sistemas CRISPR-Cas en el metagenoma generado a partir de la fracción subcelular de estas muestras de agua, primero se analizaron secuencias de

longitud ≥ 2 kb mediante el programa CRISPRCasFinder (CCFinder) con el fin de detectar genes *cas* y agrupaciones CRISPR [Couvin, D. et al. (2018). *Nucleic Acids Res.*, 46(W1), W246-W251]. A continuación, se predijeron los marcos de lectura abierta (ORF) de los 745 cóntigos con componentes CRISPR-Cas así identificados, para lo que se utilizó Prodigal v2.6.3 [Hyatt, D. et al. (2010). *BMC bioinformatics*, 11(1), 1-11]. El catálogo resultante de secuencias de proteínas se analizó con perfiles Hidden Markov Models (HMM) de dominios de proteínas Cas9 utilizando el programa hmmersearch del paquete HMMER v3.2 [Finn, R. D. et al. (2011). *Nucleic Acids Res.*, 39, W29-W37].

Como primer paso para la identificación de posibles regiones codificantes de *tracrRNA*, se buscaron secuencias semejantes a repeticiones (repeticiones degeneradas) en las proximidades de los loci CRISPR-*cas* con la plataforma en línea Benchling (<https://benchling.com/editor>). Seguidamente, se predijeron secuencias promotoras y terminadoras a ambos lados de las repeticiones degeneradas encontradas con BPROM y FindTerm [Salamov, V. S. A., & Solovyev, A. (2011). *Metagenomics and its applications in agriculture, biomedicine and environmental studies*, Nova Science Publishers, 61-78], respectivamente. Finalmente seleccionamos para su posterior análisis funcional y bioquímico un sistema (sistema CRISPR-EHCas9) asociado a un gen *cas9* (*ehcas9*) y un posible *tracrRNA*.

La Figura 1 muestra una representación esquemática del locus CRISPR-EHCas9 y los dominios de la proteína EHCas9 asociada. El locus CRISPR-EHCas9 comprende tres genes *cas*, en el orden *cas9* (denominado *ehcas9*) - *cas1* - *cas2* (representados mediante rectángulos apuntando en el sentido de la transcripción), y una agrupación EHCRISPR que consta de dos repeticiones de 36 pb con secuencia SEQ ID NO: 71 (rectángulos blancos) separadas por un espaciador de 29 pb (rombo). Corriente arriba de *ehcas9* se identificó un posible gen de *tracrRNA* (representado con una flecha que apunta en el sentido de la transcripción en la Figura 1) como una región de ~100 pb, flaqueada por un promotor y un terminador independiente de Rho, que contiene una secuencia anti-repetición (parcialmente complementaria a las unidades CRISPR asociadas). El gen *ehcas9* codifica la proteína EHCas9, cuya estructura comprende los siguientes dominios: RuvC (motivos I, II y III), Bridge Helix (BH), de reconocimiento (REC), nucleasa HNH, Phosphate Lock Loop (PLL), WED y de interacción con PAM (PI).

La comparación mediante la herramienta BLASTp de la secuencia aminoacídica de EHCas9 con las de las proteínas Cas9 disponibles en la base de datos de secuencias del

NCBI (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>), muestra una identidad de secuencia inferior al 68%. La comparación con ortólogos nativos Cas9 empleados para la edición de genomas en células de mamíferos muestra una identidad de secuencia inferior al 29% (ver Tabla 4). En concreto, estos alineamientos revelaron la arquitectura de dominios típica de esta familia de proteínas [Jinek, M. et al. (2014). *Science*, 343(6176), 1247997; Yamada, M. et al. (2017). *Mol. Cell*, 65(6), 1109-1121; Hirano, S. et al. (2019). *Nat. Commun.*, 10(1), 1-11; Nishimasu, H. et al. (2015). *Cell*, 162(5), 1113-1126; Fuchsbauer, O. et al. (2019). *Mol. Cell*, 76(6), 922-937; Sun, W. et al. (2019). *Mol. Cell*, 76(6), 938-952] con residuos catalíticos conservados en los dominios nucleasa RuvC (D11, E521, H747 y D750) y HNH (D605, H606 y N629) (ver Figuras 2A y 2B). Sin embargo, la secuencia del dominio que interactúa con PAM difiere considerablemente. En conjunto, estas observaciones sugieren que EHCas9 podría actuar como una nucleasa guiada por crRNA:tracrRNA de manera similar a los ortólogos caracterizados bioquímicamente, pero reconociendo PAM distintos.

Con respecto a su tamaño, EHCas9 está en el rango de los ortólogos más pequeños, presentando una longitud total de 1.070 aa y una masa de aproximadamente 120 kDa. Ello permite la administración de las secuencias codificantes de la herramienta EHCas9 a células eucariotas empleando un vector de tamaño restringido, como los virus adenoasociados (AAV), comúnmente utilizados en biomedicina. Además, su pequeño tamaño también puede facilitar la administración de derivados inactivos de la nucleasa fusionados con péptidos con distintas actividades relacionadas con el ADN, como se ha hecho con las proteínas dead-Cas9 (dCas9).

La relación evolutiva de EHCas9 se analizó mediante la reconstrucción de un árbol filogenético que incluye 798 secuencias de proteínas Cas9 ortólogas (Figura 3). En concreto, se llevó a cabo un alineamiento múltiple entre SEQ ID NO: 1 y las secuencias de una base de datos de ortólogos de Cas9 recopiladas por Gasiunas et al. [*Nat. Commun.* 2020, 11(1), 1-10] mediante el programa MUSCLE. El árbol filogenético se generó a partir de los alineamientos con el programa Fast Tree utilizando un modelo evolutivo JTT y un modelo gamma discreto, obteniéndose como conclusión que la proteína EHCas9 pertenece al clado IX del subtipo II-C y está lejanamente emparentada con las proteínas Cas9 comúnmente utilizadas en la edición de genomas, siendo Cas9 de *S. donghicola* (SdoCas9) la más estrechamente relacionada de entre los ortólogos caracterizados bioquímicamente.

Tabla 4. Características relevantes de las proteínas Cas9 nativas empleadas en edición de células de mamíferos.

Nombre	Subtipo CRISPR-Cas	Origen	Tamaño (aa)	PAM	Identidad (%)	Ref.
EHCas9	II-C	Metagenoma de El Hondo	1.070	NGG	100	Inención
CjCas9	II-C	CJE	984	NNNNRYAC	25,1	[8]
FnCas9	II-B	<i>F. tularensis</i> subsp. novicida U112	1.629	NGG	20,6	[9]
GeoCas9	II-C	<i>G. stearothermophilus</i>	1.087	NNNNCRAA	28,3	[10]
Nme2Cas9	II-C	<i>Nme</i> De11444	1.082	NNNNCC	26,9	[11]
Nm1Cas9	II-C	<i>Nme</i> 8013	1.082	NNNNGATT	27,1	[12]
PpCas9	II-C	<i>P. pneumotropica</i>	1.055	NNNNRTT	26,3	[13]
SaCas9	II-A	<i>S. aureus</i>	1.053	NNGRRT	23,1	[14]
SauriCas9	II-A	<i>S. auricularis</i>	1.061	NNGG	22,4	[15]
ScCas9	II-A	<i>S. canis</i>	1.375	NNG	22,3	[16]
SpCas9	II-A	<i>S. pyogenes</i> SF370	1.368	NGG	23,1	[17, 18]
St1Cas9	II-A	<i>S. thermophilus</i> LMD-9	1.121	NNAGAAW	24,6	[19]
St3Cas9	II-A	<i>S. thermophilus</i> LMD-9	1.388	NGGNG	23,9	[20]

[8] Kim, E. et al. (2017). Nat. Commun., 8(1), 1-12; [9] Hirano, H. et al. (2016). Cell, 164(5), 950-961; [10] Harrington, L. B. et al. (2017). Nat. Commun., 8(1), 1-8; [11] Edraki, A. et al. (2019). Mol. Cell, 73(4), 714-726; [12] Esvelt, K. M. et al. (2013). Nat. Methods, 10(11), 1116-1121; [13] Fedorova, I. et al. (2020). Nucleic Acids Res., 48(21), 12297-12309; [14] Ran, F. et al. (2015). Nature, 520(7546), 186-191; [15] Hu, Z. et al. (2020). PLoS Biol., 18, 1-18; [16] Chatterjee, P. et al. (2018). Sci. Adv., 4(10), eaau0766; [17] Cong, L. et al. (2013). Science, 339(6121), 819-823; [18] Mali, P. et al. (2013). Science, 339(6121), 823-826; [19] Esvelt, K. M. et al. (2013). Nat. Methods, 10(11), 1116-1121. [20] Müller, M. et al. (2016). Mol., 24(3), 636-644.

Ejemplo 2: Detección y validación *in vivo* de PAM y determinación de los requisitos del ARN guía para la escisión de ADN mediada por EHCas9.

Para el cribado *in vivo* de los motivos PAM reconocidos por EHCas9, se generó en primer lugar el plásmido pMML01 (control negativo de la actividad de EHCas9) mediante la clonación en pBAD33 de una agrupación EH CRISPR formada por dos repeticiones de 36 pb de longitud separadas por un espaciador de 29 pb. Se diseñó otro plásmido derivado de pBAD33 que lleva además el gen *ehcas9* (pMML02, Figura 4A), y un plásmido basado en pUC57 que contiene un inserto de 300 pb de longitud que abarca la secuencia codificante de EH tracrRNA (pMML03, Figura 4B). Para construir pMML02, se clonaron en el vector un gen *ehcas9* cuyos codones fueron optimizados para su expresión en *E. coli* (SEQ ID NO:4) bajo promotores inducibles, junto con una agrupación CRISPR formada por un espaciador flaqueado por dos repeticiones, transcrita a partir de un promotor constitutivo (Part:BBa_J23101, colección BioBricks) y la secuencia terminadora BBa_B1006, adquirida como bloques G de NZYTech, de tal forma que *ehcas9* queda bajo el control del promotor de arabinosa PBAD. Para la construcción de pMML03, el inserto sintetizado por NZYtech como un bloque G se clonó bajo el promotor T7 (inducible por lac/IPTG) del vector.

Por otro lado, se generó una genoteca de plásmidos derivados de pSEVA431 (resistencia a la espectinomicina) con motivos PAM aleatorios de 3 nt, mediante mutagénesis por PCR con cebadores (SEQ ID NO: 17 y SEQ ID NO: 18; ver Tabla 3) que contenían nucleótidos aleatorios en las posiciones 2, 3 y 4 respecto al extremo 3' de la secuencia diana en la hebra que coincide con el espaciador, SEQ ID NO: 72; es decir, la región PAM. En concreto, dada la tolerancia de cualquier nucleótido en la primera posición de la región PAM exhibida por la mayoría de las proteínas Cas9 [Gasiunas, G. et al. (2020). Nat. Commun., 11(1), 1-10], se mantuvo invariable una timina en esta ubicación y nucleótidos aleatorios en las posiciones 2, 3 y 4 (consenso 5'-TNNN-3').

Células electrocompetentes de *E. coli* BW 27783 fueron co-transformadas con pMML03 y con pMML01 o pMML02 y se seleccionaron en placas de LB agar conteniendo ampicilina y cloranfenicol. Las colonias transformantes se crecieron en medio líquido suplementado con ampicilina, cloranfenicol, L-arabinosa (0,2 %) e IPTG (1 mM). A continuación, se prepararon células electrocompetentes a partir de cultivos a $OD_{600} = 0,5$ y se realizaron tres experimentos de transformación independientes con 300 ng de la genoteca PAM para los portadores de pMML01 y pMML02. Los transformantes que portaban plásmidos

derivados de pSEVA431 se seleccionaron en LB agar suplementado con espectinomicina y los plásmidos fueron aislados a partir de ca. 10^5 colonias. La región del plásmido que flanquea al PAM se amplificó por PCR utilizando como cebadores SEQ ID NO: 24 y SEQ ID NO: 25 (ver Tabla 3), y se secuenció mediante secuenciación masiva (HTS) con el sistema de secuenciación Illumina NovaSeq PE250 (Novogene). La proporción de lecturas con cada secuencia PAM específica obtenida de las células portadoras de pMML02 se comparó con los valores correspondientes a las células que portaban el control negativo pMML01 para estimar su cambio log₂ (log₂ fold change). Se utilizaron secuencias PAM con un valor de log₂ superior a 7 para generar logotipos de secuencias con la aplicación WebLogo (<https://weblogo.berkeley.edu/logo.cgi>).

La comparación de la incidencia de cada secuencia en la región PAM en presencia o ausencia de EHCas9 reveló que el nucleótido guanina estaba infrarrepresentado en las posiciones 2 y 3 cuando se producía la proteína (Figura 5A), pero no se observaron diferencias en la frecuencia de cualquier nucleótido específico en la posición 4. Estos resultados demuestran que EHCas9 puede interferir de manera específica con plásmidos diana si hay una guanina en la segunda y la tercera posición del PAM. También apoyan la identidad del EH tracrRNA, así como la dirección de transcripción de la agrupación CRISPR inferida. Además, prueban que, bajo las condiciones ensayadas en *E. coli*, se genera un crRNA funcional a partir del EH pre-crRNA diseñado.

Para la implementación de una herramienta EHCas9 simplificada, se dedujo la secuencia de un sgRNA (EH sgRNA) a partir de la secuencia validada bioquímicamente del sistema tipo II-C de *S. donghicola*. Después de comparar el crRNA y el tracrRNA de los dos sistemas, se concibió un EH sgRNA de 118 nt de longitud, compuesto por una región espaciadora variable de 23 nt y una secuencia constante (SEQ ID NO:3) de 95 nt que consta de una repetición truncada de 18 nt, un conector de 4 nt (tetraloop 5'-GAAA-3') y un fragmento del EH tracrRNA de 73 nt que contiene la anti-repetición seguida de una secuencia que presumiblemente adopta dos estructuras de tallo-bucle (Figura 6).

Para probar la funcionalidad del EH sgRNA y expandir el PAM inferido a partir del cribado *in vivo*, se testaron las primeras siete posiciones de la región PAM utilizando un procedimiento de traducción *in vitro* (IVT) siguiendo el mismo procedimiento empleado previamente por otros autores [Gasiunas, G. et al. (2020). Nat. Commun., 11(1), 1-10]. Esta detección de PAM se llevó a cabo en colaboración con la empresa CasZyme, utilizando EHCas9 y un EH sgRNA con diana en una genoteca de plásmidos con

secuencias aleatorias en cada una de las 7 posiciones PAM a testar (Tabla 5). Se incluyó $MgCl_2$ en la reacción, ya que se ha demostrado que las proteínas Cas9 requieren cationes divalentes para adoptar el estado competente para la escisión [Jinek, M. et al. (2012). Science, 337(6096), 816-821; Mougiakos, I. et al. (2017). Nat. Commun., 8(1), 1-11; Chen, H. et al. (2014). J. Biol. Chem., 289(19), 13284-13294; Dagdas, Y. S. et al. (2017). Sci. Adv., 3(8), eaao0027]. El análisis de secuencias reveló la escisión de la diana, lo que corrobora la funcionalidad del EH sgRNA diseñado. Al igual que algunas nucleasas Cas9 previamente caracterizadas [Jinek, M. et al. (2012). Science, 337(6096), 816-821; Gasiunas, G. et al. (2020). Nat. Commun., 11(1), 1-10], la escisión se observó preferentemente entre los nucleótidos en las posiciones 3 y 4 respecto del PAM, en ambas cadenas de la diana, lo que sugiere la formación de extremos romos. El análisis de la región PAM (Figura 5B) confirmó que, de acuerdo con los resultados de la detección de PAM *in vivo*, guanina en las posiciones 2 y 3 es indispensable para la escisión. Sin embargo, en contraste con la tolerancia de cualquier nucleótido en la posición 4 observada *in vivo*, se evidenció una cierta discriminación contra citosina. Además, aunque para la actividad de EHCas9 no se requerían nucleótidos específicos en las posiciones restantes, se reveló una preferencia por timina en la 5ª posición, lo que sugiere que la ausencia de este nucleótido en el cribado *in vivo* podría haber comprometido el reconocimiento de la diana cuando citosina está presente en la 4ª posición. En resumen, mientras que los PAM compatibles con la escisión de la diana de EHCas9 en las condiciones *in vitro* utilizadas se corresponden con la secuencia consenso 5'-NGGNNNN-3', el PAM responde al consenso 5'-NGGDTNN-3' (D = A o T o G).

Tabla 5. Secuencias diana de Cas9 utilizadas en la validación *in vivo* de PAM.

Secuencia de la hebra coincidente con el espaciador (5' → 3') [SEQ ID NO: 72]	Región PAM (5' → 3')
CCTGTATATCGTGCGAAAAAGGATGGATA	TACCGAA
CCTGTATATCGTGCGAAAAAGGATGGATA	TGGAGAA
CCTGTATATCGTGCGAAAAAGGATGGATA	TGGCGAA
CCTGTATATCGTGCGAAAAAGGATGGATA	TGGGGAA
CCTGTATATCGTGCGAAAAAGGATGGATA	TGGTGAA

A continuación, se verificó la tolerancia por citosina en la cuarta posición del PAM junto con el requerimiento de timina en la quinta posición. Con este fin, se llevaron a cabo ensayos de transformación con plásmidos equivalentes a los utilizados para la detección de PAM *in vivo*, pero en lugar de una genoteca de PAM se emplearon plásmidos

individuales (pMML04-07; ver Tabla 2) que contienen en este caso la secuencia diana adyacente a 5'-TGGCG-3', 5'- TGGTG-3', 5'-TGGAG-3' o 5'-TGGGG-3' (Tabla 5). De la misma manera, se analizó el motivo 5'-TACCG-3' como control en ausencia de PAM. Como era de esperar, cuando el plásmido diana con la secuencia flanqueante 5'-TACCG-3' se transformó en células que expresaban los tres componentes del locus CRISPR-EHCas9, la eficiencia de la transformación no difirió significativamente de la eficiencia observada en ausencia de EHCas9. Sin embargo, se encontró una marcada disminución en la eficiencia de la transformación cuando los plásmidos 5'-TGGNG-3' se transformaron en células que expresan EHCas9 en comparación con hospedadores sin la nucleasa, mostrando una diferencia de aproximadamente cuatro órdenes de magnitud en el caso del plásmido con citosina en la 4ª posición del PAM, y aproximadamente cinco órdenes de magnitud para el resto (Figura 5C). Estos resultados confirman que, incluso en ausencia de timina en la 5ª posición, EHCas9 cataliza de forma eficaz la escisión de la diana en *E. coli* independientemente de la identidad del nucleótido en la 4ª posición, siendo citosina la que muestra una menor actividad.

Ejemplo 3: Purificación de la proteína de la presente invención con secuencia aminoacídica según SEQ ID NO: 5.

Para la expresión heteróloga de la proteína de la presente invención con secuencia aminoacídica según SEQ ID NO: 5, el gen *ehcas9* con uso de codones optimizado para *E. coli* (suministrado por NZYtech) se fusionó con una cola N-terminal de seis histidinas (SEQ ID NO: 6) bajo un promotor inducible por lac/IPTG en un vector pHTP1, generando el plásmido pMML22 (Figura 4F). *E. coli* BL21(DE3) previamente transformada con pMML22 se creció a +37°C en LB suplementado con kanamicina. Cuando el cultivo alcanzó una DO₆₀₀ = 0,5, se indujo la expresión de la proteína añadiendo IPTG 1 mM y tras 16 h de incubación a +16°C se recolectaron las células por centrifugación (5.000 x g durante 15 min a +4°C) y se resuspendieron en tampón de unión compuesto por tampón fosfato pH 7,6 (50 mM), NaCl (500 mM), imidazol (10 mM), glicerol (5%), β-mercaptoetanol (10 mM) y fluoruro de fenilmetilsulfonilo (PMSF; 1 mM). Las células se rompieron mediante sonicación con un Branson Digital Sonifier®. Después de centrifugar (23.700 x g durante 25 min a +4°C), el sobrenadante se cargó en una columna HisTrap HP de 1 ml (GE Healthcare), se lavó la columna con 20 volúmenes de tampón de unión y se la proteína se eluyó con tampón de elución (tampón fosfato 50 mM pH 7,6, 500 NaCl mM, 150 mM imidazol, glicerol al 5%, 10 mM β-mercaptoetanol, 1 mM PMFS). La fracción eluida se concentró hasta un volumen de 1 ml en tampón de digestión (tampón fosfato 50 mM pH

7,6, 150 mM NaCl, glicerol al 5%, 10 mM β -mercaptoetanol) utilizando filtros Amicon Ultra (Millipore) y se cargó en un gel de filtración HiLoad™ 16 /600 Superdex™ 200 pg (Cytiva). Las fracciones eluidas se analizaron mediante SDS-PAGE y la fracción conteniendo una proteína del tamaño esperado para EHCas9 se concentró como se indica anteriormente (Figura 7).

Para la estimación del tamaño de las proteínas se utilizó NZYBlue Protein Marker (NZYtech) y la concentración de proteínas se midió con QUBIT® 2.0 (Invitrogen).

10 Ejemplo 4: Optimización *in vitro* de las condiciones de reacción requeridas para la escisión de diana mediada por la proteína de la presente invención.

Para la optimización de las condiciones de reacción requeridas por la proteína de la presente invención para la escisión de ADN bicatenario (ADNbc), se diseñó y generó *in vitro* un EH sgRNA. Para obtener un molde de ADNbc mediante amplificación por PCR de la región constante codificante de sgRNA (SEQ ID NO:3) a partir del plásmido pMML08, se utilizaron oligonucleótidos que portan un promotor T7 y una secuencia coincidente con el espaciador de 23 nt de longitud en pSEVA431 (SEQ ID NO: 26 y SEQ ID NO: 27; ver Tabla 3). El amplicón se transcribió con HiScribe T7 Quick (NEB) siguiendo las instrucciones del fabricante, incluido el tratamiento opcional con ADNasa, y el ARN se purificó con el kit de limpieza de ARN Monarch® (NEB). Las alícuotas de sgRNA se almacenaron a -80°C.

Como sustrato de escisión se utilizó un fragmento de 840 pb amplificado por PCR a partir de pMML05 (derivado de pSEVA431 que contiene una diana con PAM 5'-TGGCG-3'). Como control sin PAM, se amplificó un fragmento de pMML05 que contenía una diana con la secuencia 5'-TACCG-3' en la región de la PAM (Tabla 6). La escisión específica de la diana guiada por EH sgRNA producirá dos fragmentos de ADNbc (520 pb y 320 pb de longitud, respectivamente).

Tabla 6. Secuencias diana de Cas9 utilizadas en la optimización *in vitro* de las condiciones de reacción requeridas para la escisión de diana mediada por EHCas9.

Experimento	Secuencia de la hebra coincidente con el espaciador (5' → 3')	Región PAM (5' → 3')
Cribado <i>in vitro</i> de PAM	SEQ ID NO: 73	NNNNNNN

Actividad de escisión <i>in vitro</i>	SEQ ID NO: 74	TACCGAA
Actividad de escisión <i>in vitro</i>	SEQ ID NO: 75	TGGCGAA

En primer lugar, evaluamos la especificidad de escisión de la diana de ADNbc a +37°C y el requerimiento de Mg²⁺ (Figura 8A). Para facilitar la formación del complejo ribonucleoproteico, pre-incubamos (15 min a +37°C) la nucleasa con EH sgRNA (relación molar 1:1) antes de mezclarla con la diana (la relación molar final Cas9:sgRNA:diana en la solución de reacción fue 20:20:1) en presencia de MgCl₂. Como era de esperar, la pre-incubación aumentó la tasa de escisión de la diana en comparación con reacciones en las que todos los componentes se mezclaron simultáneamente (30 minutos después de añadir al sustrato la proteína pre-incubada con la guía o ambas soluciones sin pre-incubar, se había escindido el 21,6% y el 15,6% de sustrato, respectivamente). En base a estos resultados, los experimentos *in vitro* posteriores con EHCas9 y EH sgRNA se llevaron a cabo después de pre-incubar en las mismas condiciones ensayadas. No se observaron productos de escisión en la diana sin PAM, ni cuando no se agregó EH sgRNA o Mg²⁺ a la reacción. En presencia de todos los reactivos, el sustrato con el PAM compatible se cortó una vez, generando dos fragmentos de ADN cuyos tamaños coincidían con los esperados por la escisión dentro de la secuencia diana. Estos resultados corroboran que EHCas9 es una endonucleasa de ADNbc dependiente de metales, específica de secuencia y guiada por ARN.

Seguidamente, se caracterizó la actividad de corte de ADNbc guiada por ARN en presencia de MgCl₂, bajo distintos tiempos de digestión y temperatura. Para decidir la cantidad de EHCas9 a utilizar en estos experimentos, se pre-incubaron concentraciones constantes de EH sgRNA durante 15 min a +37°C con 10 nM a 0,5 µM de EHCas9 y posteriormente se mezclaron con una concentración fija de sustrato, de modo que la relación molar proteína:sgRNA:sustrato en la reacción de digestión varió de 1:50:2,5 a 20:20:1. Concentraciones de proteína por encima de 0,1 µM produjeron productos de digestión perceptibles después de 30 min, eligiéndose una concentración de EHCas9 de 0,5 µM para los posteriores ensayos de temperatura y tiempo de incubación (Figura 8B). Cuando se evaluaron diferentes tiempos de reacción (hasta 40 min) a +37°C, aunque se cortó una proporción sustancial (21,6%) del sustrato dentro de los primeros 5 minutos, lo que subraya la robustez de la nucleasa, el máximo porcentaje de digestión (alrededor del 27% de sustrato escindido) se alcanzó después de 30 min (Figura 8C). Curiosamente, la incubación durante 10 minutos más no aumentó la cantidad de sustrato cortado, lo que sugiere que EHCas9 permanece unido al ADN después de catalizar su escisión, evitando

así que actúe sobre otras moléculas diana. En cuanto a la temperatura de incubación, en los ensayos de digestión realizados a intervalos de +5°C dentro del rango +20 a +45°C, únicamente se detectaron productos de digestión a +30°C y +35°C, estableciéndose un rango de temperatura de trabajo entre por encima de +25°C y menos de +40°C, con temperatura óptima alrededor de +35°C (Figura 8D).

Ejemplo 5: Uso de la herramienta EHCas9 para la selección positiva de células de *E. coli* con genoma editado.

Para la selección de células de *E. coli* con genoma editado (Figura 9A), se construyó el plásmido pMML09 (Figura 4C) que codifica EHCas9 y un EH sgRNA dirigido al gen cromosómico *pyrF*, a partir de pMML02 reemplazando la región entre el promotor y el terminador de la agrupación CRISPR con una secuencia codificante de sgRNA que contiene un espaciador que coincide con una secuencia de *pyrF*, ubicado junto a la secuencia 5'-TGGAT-3' en la región PAM (SEQ ID NO: 76). Como control negativo de la actividad de EHCas9, se generó un plásmido sin *ehcas9* (pMML10) mediante amplificación por PCR de pMML09.

Mediante ensamblaje por Gibson se generó un molde de recombinación de ADN lineal de 308 pb consistente en secuencias flanqueantes de *pyrF*; en concreto, una secuencia de 145 pb que coincide con la región intergénica corriente arriba de *pyrF* y una secuencia de 163 pb que coincide con la región aguas abajo del gen.

Se transformaron células electrocompetentes de *E. coli* BW 27783 con el plásmido pKD46 (resistencia a la ampicilina) que codifica el sistema de recombinación Lambda Red (Exo, Beta, Gam) [Datsenko, K. A., & Wanner, B. L. (2000). Proc. Natl. Acad. Sci. U.S.A, 97(12), 6640-6645]. Dado que la replicación de este plásmido es sensible a la temperatura, inhibiéndose a +37°C, los transformantes se crecieron a +30°C en placas de LB agar conteniendo ampicilina. Las colonias portadoras de pKD46 se transfirieron a un medio líquido suplementado con ampicilina y se crecieron a +30°C hasta una $DO_{600} = 0,2$. A continuación, se añadió L-arabinosa al 0,2 % para inducir la expresión de las proteínas Lambda Red y cuando se alcanzó una $DO_{600} = 0,5$ se prepararon células electrocompetentes a partir del cultivo. A continuación, 3 alícuotas fueron co-transformadas con 150 ng de ADN molde y con 50 ng de pMML09 o de pMML10.

Las colonias transformantes de tres experimentos independientes se crecieron en LB agar suplementado con cloranfenicol (selección de plásmidos pMML09 y pMML10) y L-arabinosa al 0,2 % (inducción de la transcripción de *ehcas9*) a +37°C, impidiendo de esta manera la replicación de pKD46. La región *pyrF* se amplificó por PCR a partir de 90 colonias seleccionadas aleatoriamente (20 de cada experimento con el plásmido que expresa EHCas9 y 10 de cada una de las réplicas de control negativo). La electroforesis en gel de agarosa al 1% de los productos de la PCR invariablemente reveló una sola banda, cuyo tamaño se correspondía con el del fragmento deletado en el caso de los clones que expresan EHCas9 o con el de la secuencia nativa para el control negativo (Figura 9B). Estos resultados demuestran la eficacia de EHCas9 como agente antibacteriano específico de secuencia y su idoneidad como complemento para aplicaciones que se benefician de la selección positiva de mutantes de *E. coli*, incluida la edición del genoma.

Ejemplo 6: Edición genética de células de mamífero mediada por la proteína de la presente invención.

Para los ensayos de edición genética en células de mamíferos, el plásmido hCas9 (Addgene #41815; Mali, P. et al. Science, 2013, 339 (6121), 823-826) que lleva el gen *spcas9* fusionado a una secuencia de localización nuclear (SV40 NLS) controlada por un promotor constitutivo de citomegalovirus (CMV), y el plásmido MLM3636 (Addgene #43860) que codifica un sgRNA compatible (Sp sgRNA) bajo el promotor constitutivo U6, se utilizaron como base para construir plásmidos equivalentes donde las secuencias codificantes de SpCas9 y Sp sgRNA fueron reemplazadas por el gen *ehcas9* con uso de codones optimizado para humanos (pMML12, Figura 4D) y una región constante de EH sgRNA (pMML13, Figura 4E), respectivamente (Fig. 10A). De este modo, el plásmido pMML12 lleva el gen *ehcas9* fusionado a una secuencia SV40 NLS (SEQ ID NO:8). Los dos insertos se adquirieron de NZYTech como bloques G.

Células Neuro-2a (N2a) de *Mus musculus* (neuroblastos de ratón; ATCC, CLC-131™) se mantuvieron en Dulbecco's Modified Eagle's Medium (DMEM) suplementado con glucosa (Sigma) y suero bovino fetal al 10%, 10 mM HEPES pH 7,4, L-glutamina 2 mM, penicilina 100 UI/ml y estreptomycin 100 µg/ml, a +37°C con 5% de CO₂ y 95% de humedad.

Se testaron cuatro regiones diana del genoma de ratón, ubicadas en los genes *Oca2* (*Oca2.2*, *Oca2.3*, *Oca2.4*) y *Lrmda* (*Lrmda.1*), adyacentes a 5'-TGGGA-3', 5'-TGGAT-3',

5'-TGGCA-3' y 5'-TGGTG-3' en la región PAM, respectivamente (Figura 10B y Tabla 7). La longitud de la región espaciadora del sgRNA es un determinante importante de la precisión del reconocimiento de dianas [Hirano, S. et al. (2019). Nat. Commun., 10(1), 1-11; Fedorova, I. et al. (2020). Nucleic Acids Res., 48(21), 12297-12309; Kim, E. et al. (2017). Nat. Commun., 8(1), 1-12; Harrington, L. B. et al. (2017). Nat. Commun., 8(1), 1-8; Edraki, A. et al. (2019). Mol. Cell, 73(4), 714-726]. Se decidió usar un espaciador con 23 nt, ya que esta longitud es efectiva en la mayoría de las proteínas Cas9 probadas previamente para la edición del genoma de mamíferos, incluida SpCas9.

10 **Tabla 7.** Secuencias diana de Cas9 utilizadas para la edición génica de células de mamífero.

Experimento	Secuencia de la hebra coincidente con el espaciador (5' → 3')	Región PAM (5' → 3')
Edición de <i>Lrmda.1</i>	SEQ ID NO: 77	TGGTGCC
Edición de <i>Oca 2.2</i>	SEQ ID NO: 78	TGGGATT
Edición de <i>Oca 2.3</i>	SEQ ID NO: 79	TGGATGT
Edición de <i>Oca 2.4</i>	SEQ ID NO: 80	TGGCAGC

En primer lugar, se evaluó la toxicidad celular de las herramientas EHCas9 y SpCas9. Soluciones de células N2a se depositaron en placas de 96 pocillos a una densidad de 1,5·10⁴ células/mL por pocillo en un volumen total de 100 µl de DMEM sin antibióticos y se co-transfectaron con 200, 150 y 100 ng de pMML12 o hCas9 y 100 ng de pMML13 o MLM3636, respectivamente. Las transfecciones se realizaron con Lipofectamine 2000 (Invitrogen), siguiendo las instrucciones del fabricante. Tres días después de la transfección, las células se fijaron con paraformaldehído al 4% durante 30 min a temperatura ambiente y, tras teñir los núcleos celulares con DAPI, se contaron con un lector de fluorescencia Spark® (TECAN) (Figura 11). Aunque se observó una ligera disminución en el número de núcleos en relación con células no transfectadas, no se encontraron diferencias significativas entre las dos herramientas Cas9. Por lo tanto, este efecto adverso sobre el crecimiento celular se consideró aceptable para proseguir con los experimentos de edición génica.

A continuación, se evaluó la aplicabilidad de EHCas9 como herramienta de edición génica mediante el análisis de inserciones y deleciones (INDELs), detectadas tras secuenciación HTS de la región diana amplificada por PCR después de co-transfectar plásmidos codificantes de EHCas9 y EH sgRNA en células N2a (Figura 10B). En concreto,

las células N2a se sembraron en placas de 24 pocillos a una densidad de $4 \cdot 10^5$ células/mL por pocillo en un volumen total de 500 µl de DMEM sin antibióticos y se co-transfectaron con 1 µg de pMML12 o hCas9 y 500 ng del plásmido codificante del sgRNA correspondiente (pMML18-pMML21 o pMML14-pMML17 respectivamente). Las transfecciones se realizaron con Lipofectamine 2000 (Invitrogen), siguiendo las instrucciones del fabricante. El ADN genómico se extrajo de las células recogidas 72 horas después de la transfección mediante el kit High Pure PCR Template Preparation (Roche). Se incluyeron controles negativos que carecían de EH sgRNA y se realizaron experimentos equivalentes con los componentes de la herramienta SpCas9.

Para el análisis de frecuencia de INDELs, se generaron amplicones de 300-400 pb mediante amplificación por PCR de las regiones que flaquean la diana, utilizando como molde 100 ng de ADN genómico de N2a. Los productos de PCR se secuenciaron en Novogene utilizando Illumina NovaSeq 6000. Las lecturas de baja calidad y los adaptadores se eliminaron con Trimmomatic v0.39 (parámetros: java -jar trimmomatic-0.39.jar PE ILLUMINACLIP:2:30:10 SLIDINGWINDOW:4:15 MINLEN:50). Las lecturas de secuenciación se contrastaron con la secuencia diana mediante el programa Bowtie2 v2.4.2 (87) y se convirtieron al formato de archivos BAM con el paquete Samtools [Li, H. et al. (2009). *Bioinformatics*, 25, 2078–2079]. El análisis de INDELs se realizó con R Core Team (2021) utilizando el paquete CrispRVariants 1.20.0 [Lindsay, H. et al. (2016) *Nat Biotechnol*, 34, 701–702]. Dicho análisis reveló INDELs para las cuatro dianas cuando se empleó la herramienta SpCas9. Con EHCas9, se detectaron INDELs alrededor del sitio diana tan solo en el caso de *Oca2.3*. Es de destacar que *Oca2.3* es la única diana ensayada con timina en la quinta posición del PAM (5'-TGGAT-3').

La eficiencia de la edición de *Oca2.3* se cuantificó como la proporción de lecturas con INDELs encontradas en esa muestra, excluyendo otras variaciones de secuencia que podrían estar presentes en la población debido a mutaciones espontáneas (Figura 10C). La herramienta EHCas9 dio lugar a un 0,84% de lecturas con la secuencia *Oca2.3* modificada, mientras que la eficiencia de edición encontrada con SpCas9 fue de 3,92%. Cabe resaltar que la identidad y la frecuencia relativa de los alelos mutados fueron similares para ambas proteínas.

REIVINDICACIONES

1. Proteína endonucleasa Cas9 que comprende una secuencia aminoacídica según SEQ
5 ID NO: 1.
2. Proteína según la reivindicación 1, que comprende una secuencia aminoacídica con al
menos un 70% de identidad de secuencia con SEQ ID NO: 1.
- 10 3. Proteína según la reivindicación 2, que comprende una secuencia aminoacídica según
SEQ ID NO: 5.
4. Proteína según la reivindicación 2, que comprende una secuencia aminoacídica según
SEQ ID NO: 7.
- 15 5. Secuencia nucleotídica que codifica la proteína según cualquiera de las
reivindicaciones 1-4.
6. Secuencia nucleotídica según la reivindicación 5, que comprende una secuencia
20 nucleotídica según SEQ ID NO: 2.
7. Secuencia nucleotídica según la reivindicación 5, que comprende una secuencia
nucleotídica según SEQ ID NO: 4.
- 25 8. Secuencia nucleotídica según la reivindicación 5, que comprende una secuencia
nucleotídica según SEQ ID NO: 6.
9. Secuencia nucleotídica según la reivindicación 5, que comprende una secuencia
nucleotídica según SEQ ID NO: 8.
- 30 10. Vector de expresión que comprende una secuencia nucleotídica según cualquiera de
las reivindicaciones 5-9.
11. Célula que comprende una proteína según cualquiera de las reivindicaciones 1-4 y/o
35 una secuencia nucleotídica según cualquiera de las reivindicaciones 5-9, y/o un vector
según la reivindicación 10.

12. Sistema CRISPR-Cas que comprende un ARN guía y una proteína según cualquiera de las reivindicaciones 1-4.
- 5 13. Sistema CRISPR-Cas según la reivindicación 12, que comprende un ARN guía procedente de la transcripción de una secuencia nucleotídica según SEQ ID NO: 3.
- 10 14. Método para editar un genoma que incluye una secuencia nucleotídica diana, donde dicho método comprende una etapa de poner en contacto una secuencia nucleotídica diana con un sistema CRISPR-Cas según cualquiera de las reivindicaciones 12-13.
- 15 15. Uso de una proteína según cualquiera de las reivindicaciones 1-4, y/o una secuencia nucleotídica según cualquiera de las reivindicaciones 5-9, y/o un vector de expresión según la reivindicación 10, y/o una célula según la reivindicación 11, y/o un sistema CRISPR-Cas según cualquiera de las reivindicaciones 12-13 para:
- la modificación genética, regulación de expresión génica y/o visualización *in vivo* de secuencias nucleotídicas concretas; y/o
 - el diagnóstico molecular de enfermedades; y/o
 - la producción de antimicrobianos específicos de secuencia.
- 20 16. Uso según la reivindicación 15 para la modificación genética, regulación de expresión génica y/o visualización *in vivo* de secuencias nucleotídicas concretas de células eucariotas.
- 25 17. Uso según la reivindicación 15 para la producción de antibacterianos.
18. Uso según la reivindicación 17 para la producción de antibacterianos frente a *Escherichia coli*.

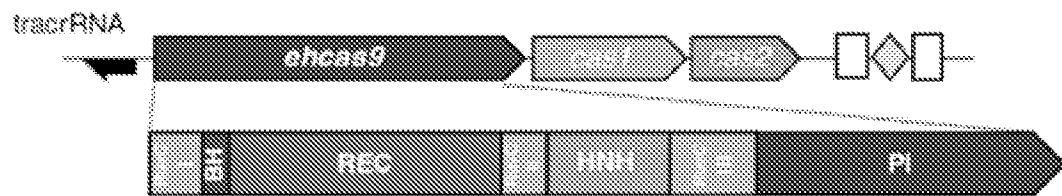
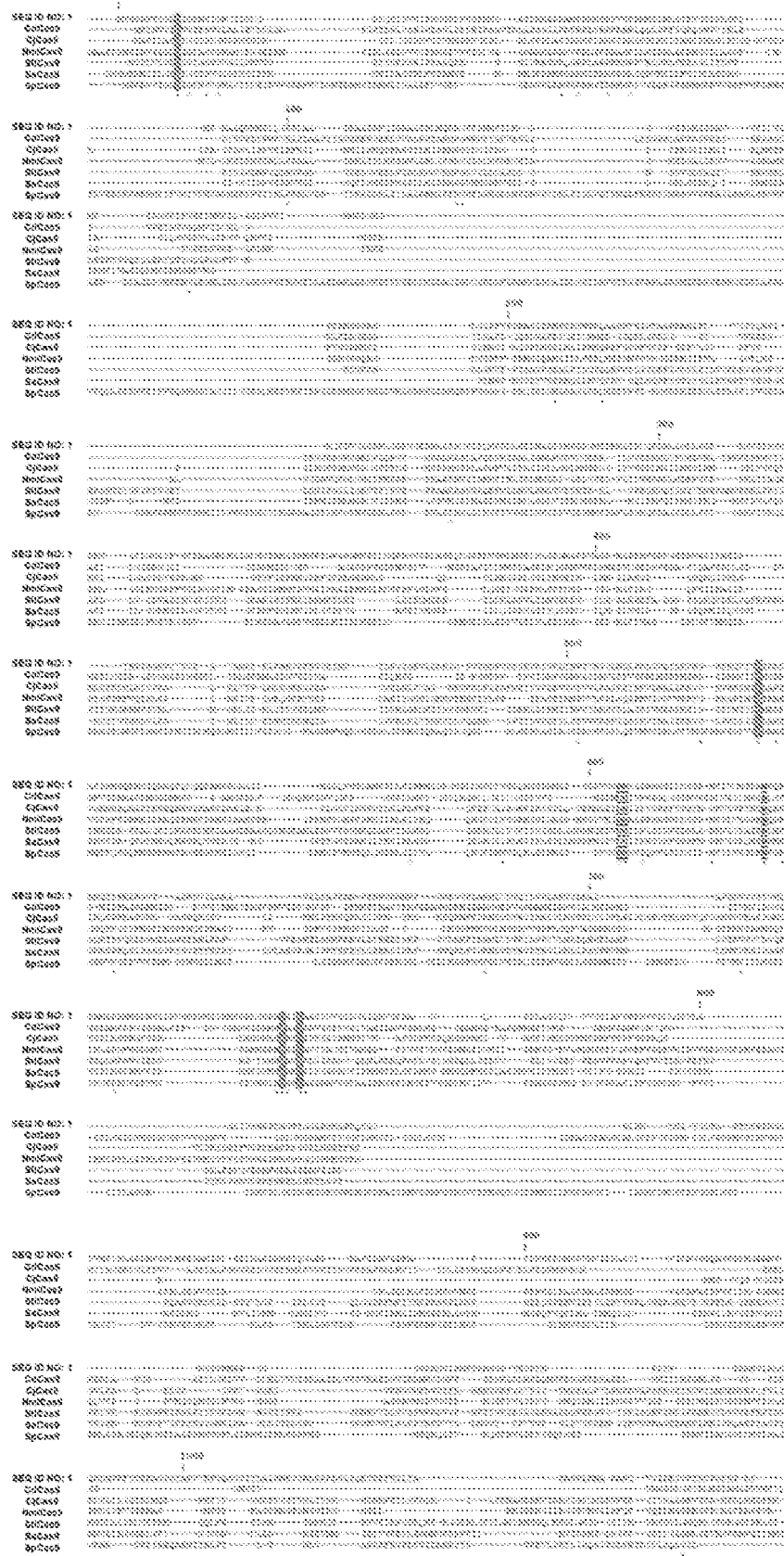


FIG. 1



FIG. 2A



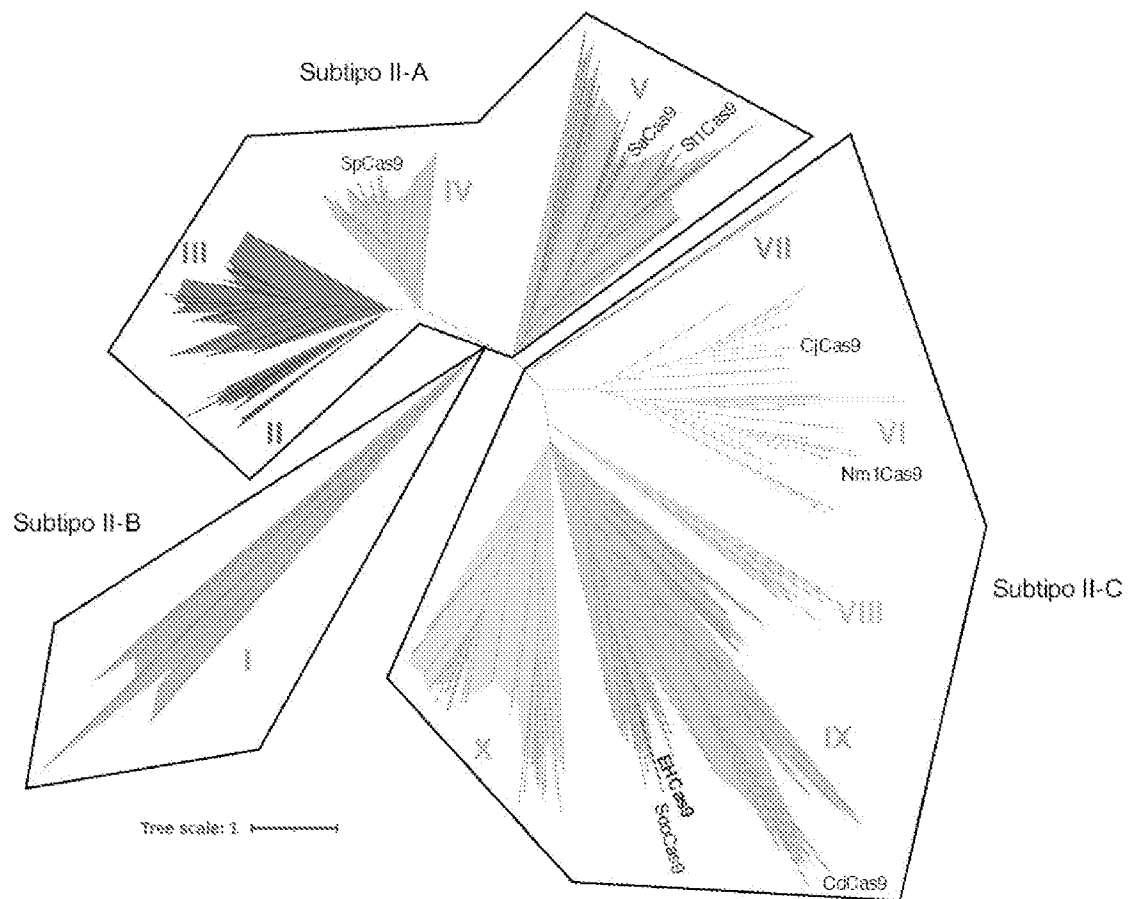


FIG. 3

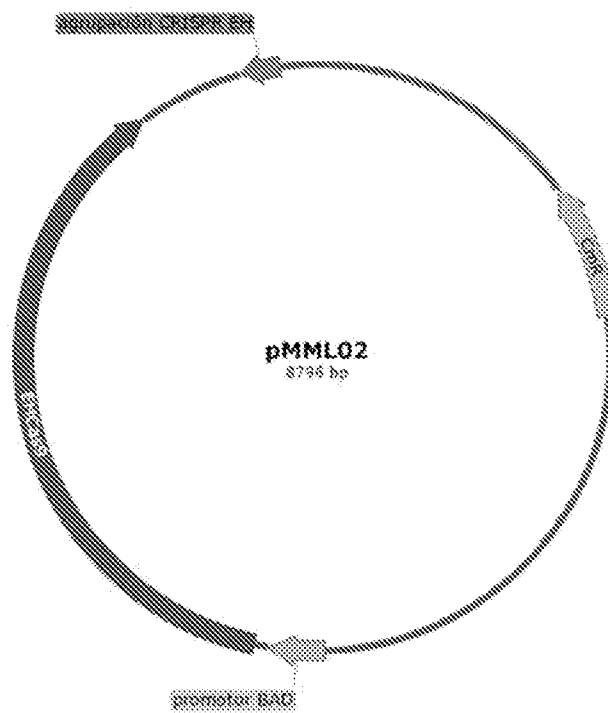


FIG. 4A

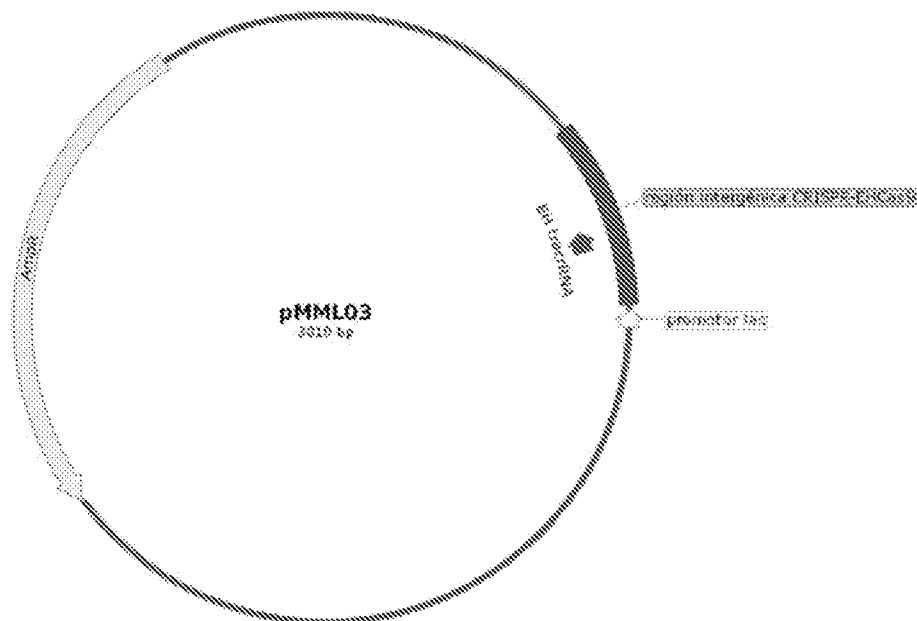


FIG. 4B

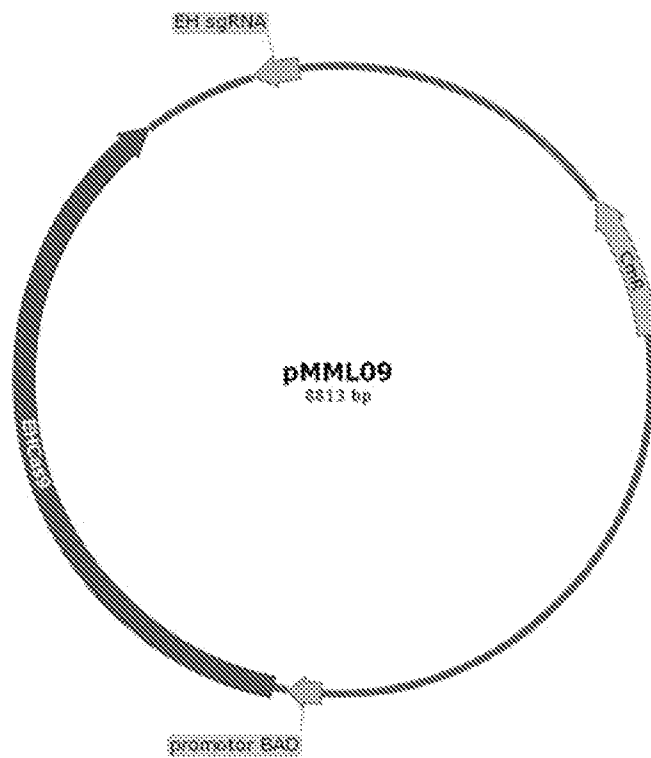


FIG. 4C



FIG. 4D

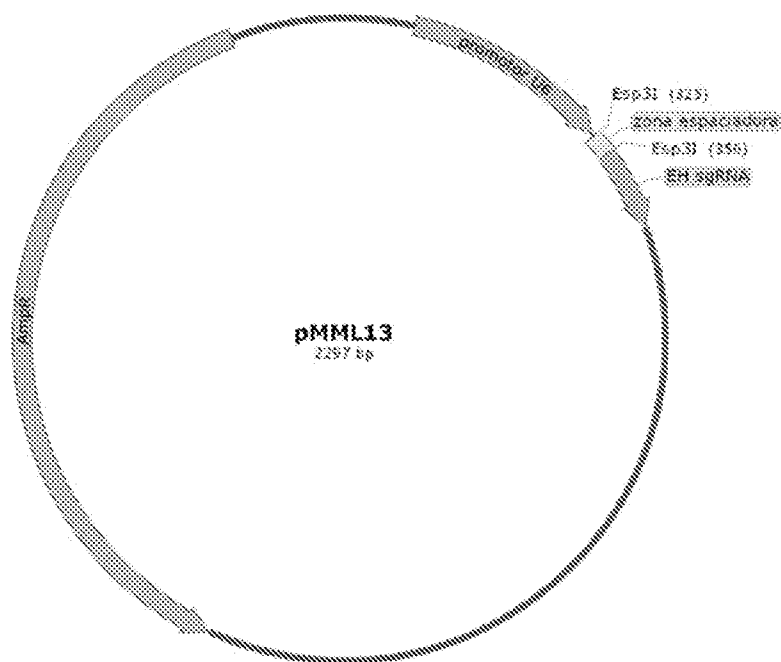


FIG. 4E

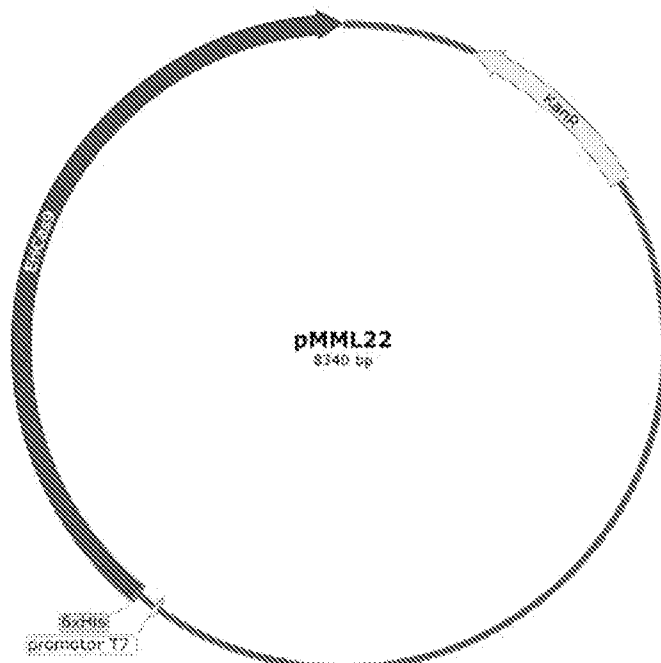
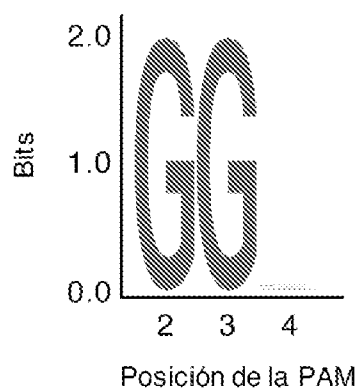
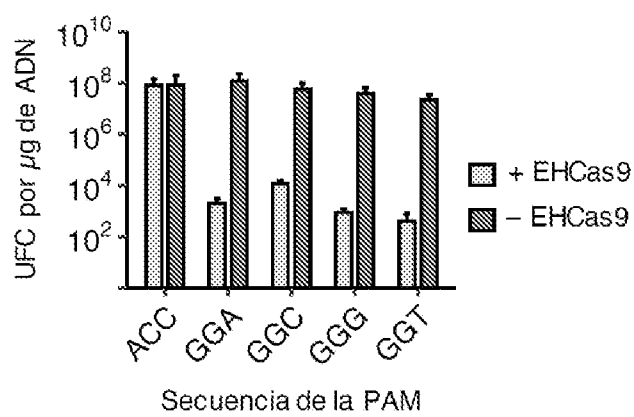
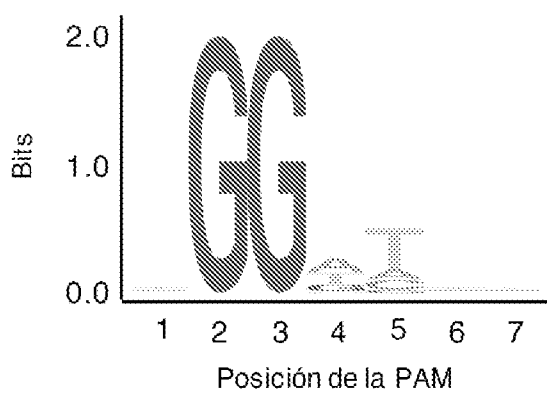


FIG. 4F

A**B****C****FIG. 5**

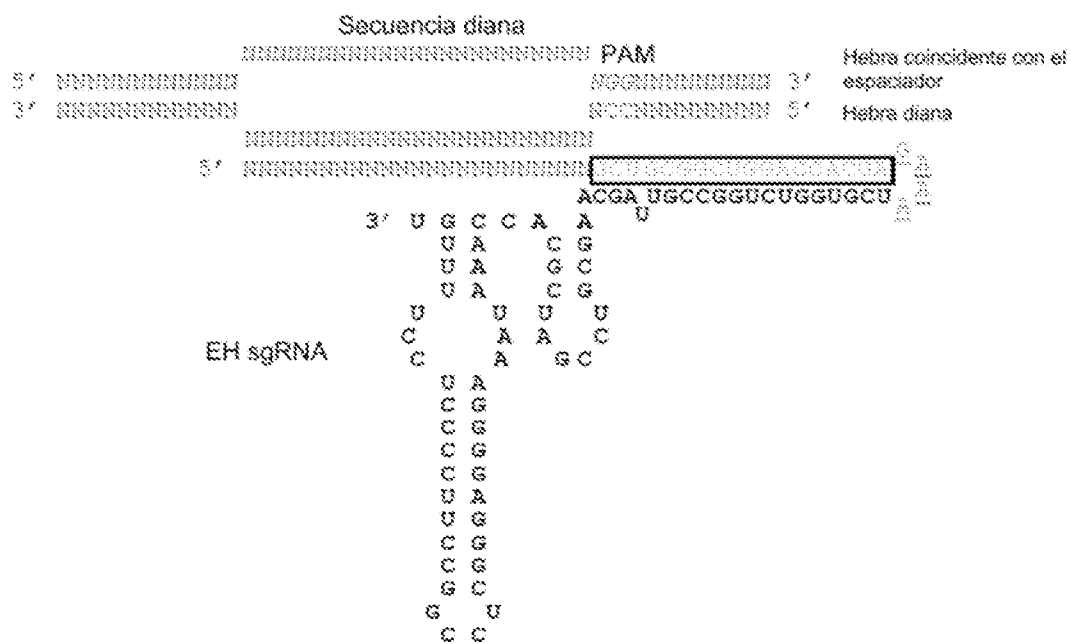


FIG. 6

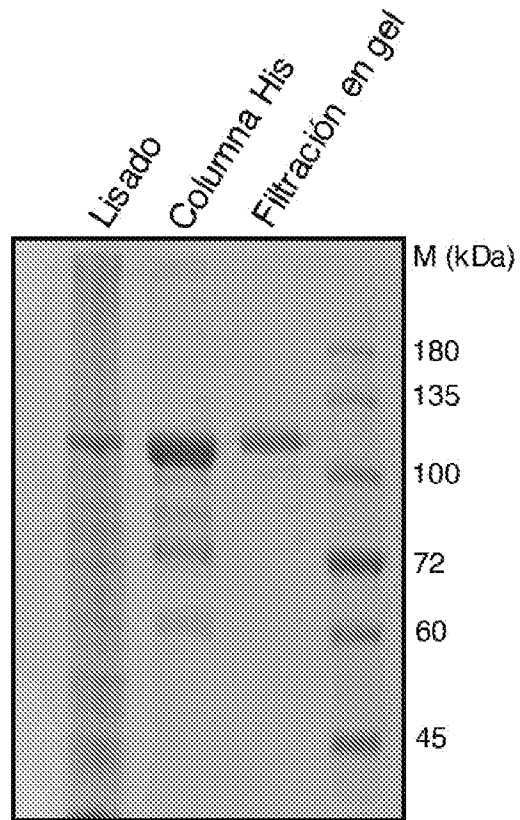
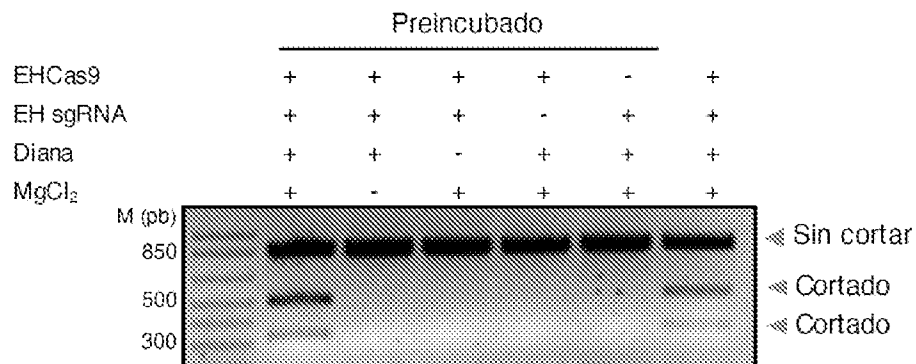
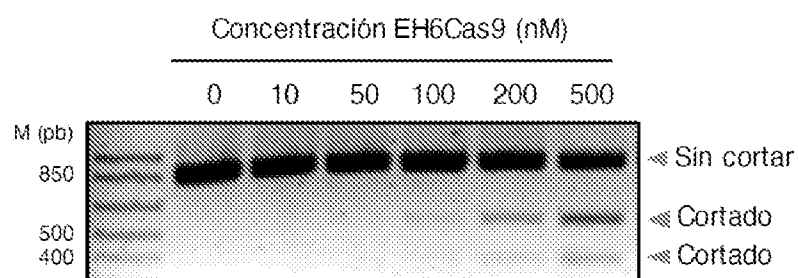


FIG. 7

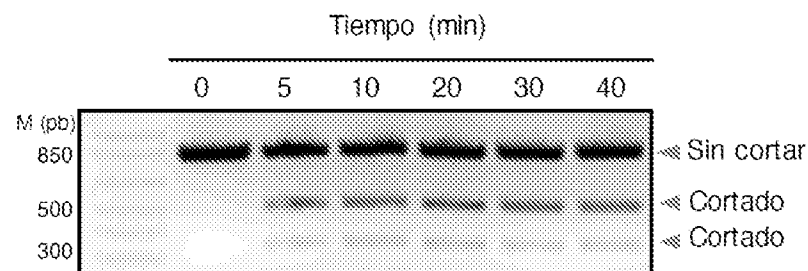
A



B



C



D

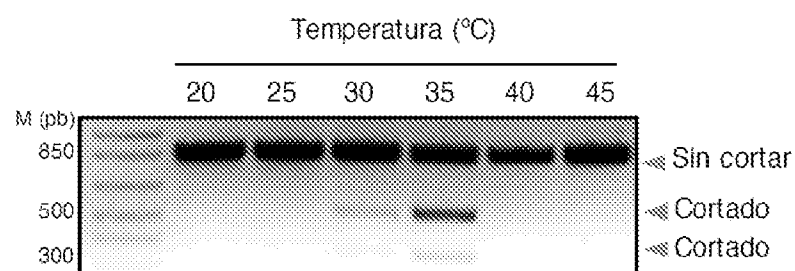
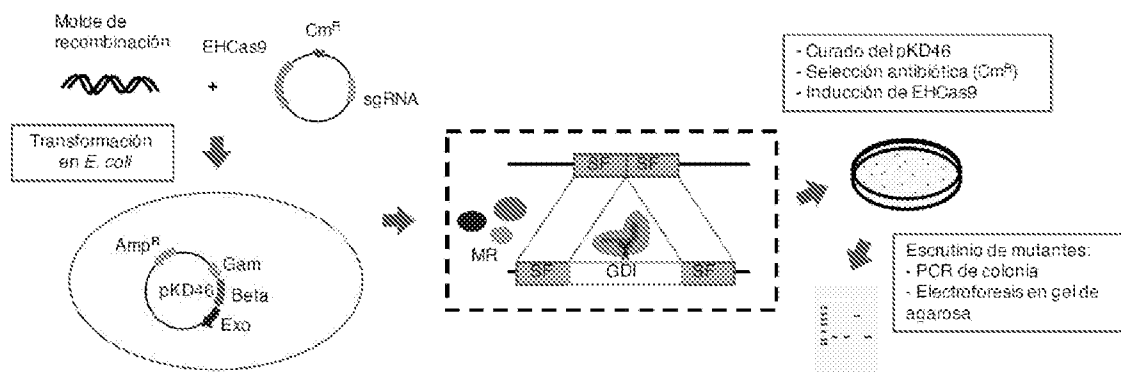


FIG. 8

A



B

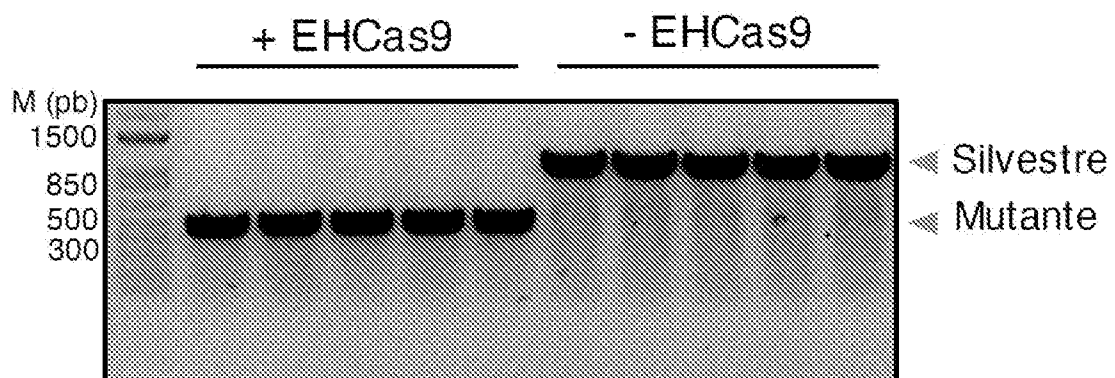
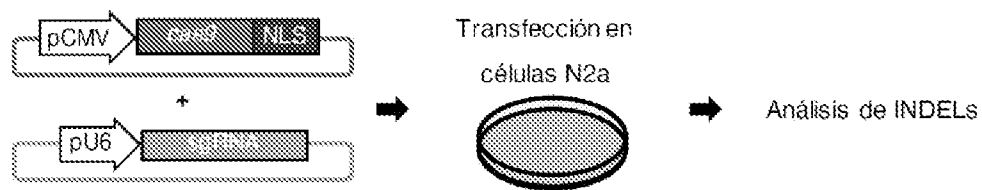
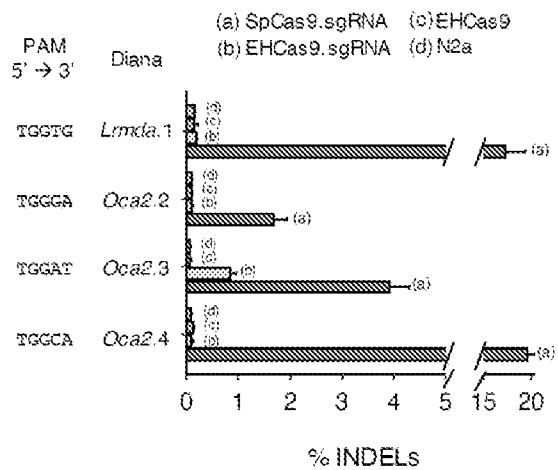
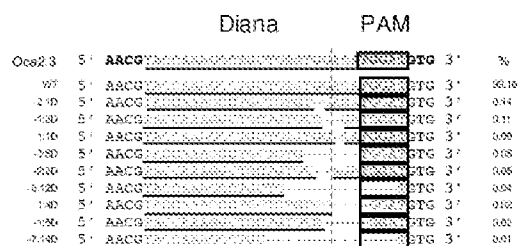
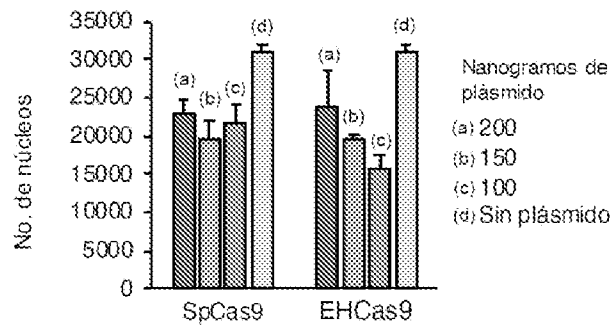


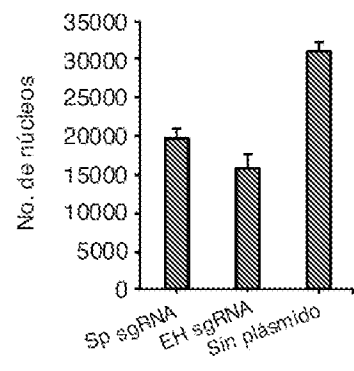
FIG. 9

A**B****C****FIG. 10**

A



B



C

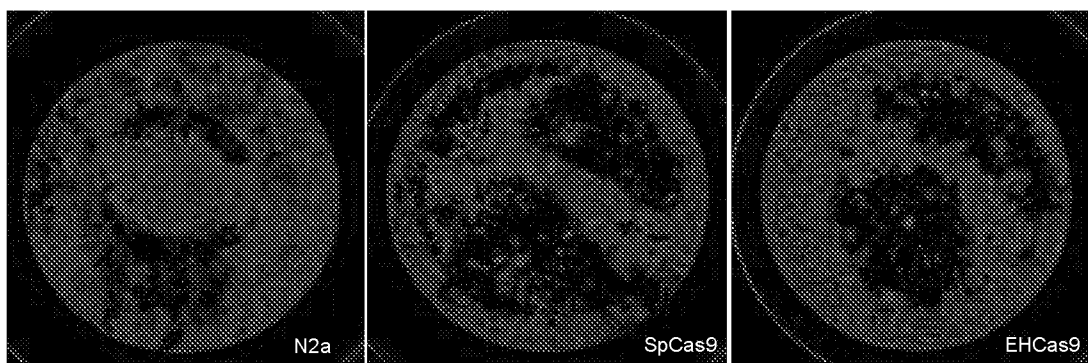


FIG. 11



OFICINA ESPAÑOLA
DE PATENTES Y MARCAS

ESPAÑA

②① N.º solicitud: 202230911

②② Fecha de presentación de la solicitud: 21.10.2022

③② Fecha de prioridad:

INFORME SOBRE EL ESTADO DE LA TECNICA

⑤ Int. cl.: **C07K14/195** (2006.01)
C12N15/63 (2006.01)

DOCUMENTOS RELEVANTES

Categoría	⑤⑥ Documentos citados	Reivindicaciones afectadas
A	KIM EUNJI et al. "In Vivo Genome Editing with a Small Cas9 Orthologue Derived from <i>Campylobacter Jejuni</i> ". Molecular Therapy (MAY 2017), 30/04/2017, Vol. 25, Nº 5, Suppl. 1, Páginas 79 [en línea] [recuperado el 29/08/2023], ISSN 1525-0016(print) ISSN 1525-0024(electronic). (ver apartados "abstract" y "results").	1-18
A	MANGHWAR HAKIM et al. "CRISPR/Cas System: Recent Advances and Future Prospects for Genome Editing". TRENDS in Plant Science, (2019) ELSEVIER, AMSTERDAM, NL, 11/11/2019, Vol. 24, Nº 12, Páginas 1102 - 1125 [en línea] [recuperado el 31/08/2023], ISSN 1360-1385, <DOI: doi: 10.1016/j.tplants.2019.09.006>. (ver Tabla 1)	1-18
A	RAGURAM ADITYA et al. "Therapeutic in vivo delivery of gene editing agents". Cell, (2022) Elsevier, Amsterdam NL, 06/07/2022, Vol. 185, Nº 15, Páginas 2806 - 2827 [en línea] [recuperado el 31/08/2023], ISSN 0092-8674, <DOI: doi: 10.1016/j.cell.2022.03.045>. (ver apartado "Development of single-AAV vectors enabled by smaller Cas orthologs")	1-18

Categoría de los documentos citados

X: de particular relevancia

Y: de particular relevancia combinado con otro/s de la misma categoría

A: refleja el estado de la técnica

O: referido a divulgación no escrita

P: publicado entre la fecha de prioridad y la de presentación de la solicitud

E: documento anterior, pero publicado después de la fecha de presentación de la solicitud

El presente informe ha sido realizado

☒ para todas las reivindicaciones

☐ para las reivindicaciones nº:

Fecha de realización del informe
31.08.2023

Examinador
M. d. García Coca

Página
1/2

Documentación mínima buscada (sistema de clasificación seguido de los símbolos de clasificación)

C07K, C12N

Bases de datos electrónicas consultadas durante la búsqueda (nombre de la base de datos y, si es posible, términos de búsqueda utilizados)

INVENES, EPODOC, WPI, BIOSIS, MEDLINE/NLM, EMBASE/Elsevier, XPESP, EMBL/EBI y Bases de Datos TXT