

DETECCIÓN DEL CÁNCER HEPÁTICO USANDO LA FRAGMENTACIÓN DEL ADN LIBRE DE CÉLULAS

La presente solicitud reivindica el beneficio de prioridad de la solicitud provisional de EE.UU.

- 5 No. 63/423.003 presentada el 6 de noviembre de 2022, que se incorpora en la presente por referencia en su totalidad.

DECLARACIÓN RELATIVA A LA INVESTIGACIÓN PATROCINADA POR EL GOBIERNO FEDERAL

- Esta invención se ha realizado con apoyo gubernamental en virtud de las subvenciones
10 GM136577, CA121113, CA006973 y CA233259 concedidas por los Institutos Nacionales de Salud. El gobierno tiene ciertos derechos sobre la invención.

ANTECEDENTES DE LA INVENCION

- El cáncer hepático es causa de una asombrosa cantidad de morbilidad y mortalidad en todo el mundo, con más de 900 mil nuevos casos diagnosticados cada año y más de 800 mil
15 muertes (1). En Estados Unidos, el cáncer hepático es uno de los pocos cánceres cuya incidencia y mortalidad han aumentado en los últimos 20 años. 90 % de los casos de cáncer hepático son carcinomas hepatocelulares (HCC, por sus siglas en inglés) y la supervivencia depende en gran medida de la etapa de la enfermedad en el momento del diagnóstico. La supervivencia a cinco años es de 34 % cuando el cáncer está localizado (44 % de los
20 pacientes), de 12 % cuando es regional (27 % de los pacientes) y de 3 % cuando se encuentra enfermedad a distancia (18 % de los pacientes) (2). Existe una población amplia y bien definida con un riesgo significativamente mayor de padecer HCC, que incluye a personas con infección crónica por el virus de la hepatitis B (HBV, por sus siglas en inglés) o con cirrosis por diversas causas, que incluye la hepatitis C (HCV, por sus siglas en inglés) (3), la
25 enfermedad del hígado graso no alcohólico (NAFLD, por sus siglas en inglés) (4), el consumo excesivo de alcohol (5), la aflatoxina y otras condiciones (6). En todo el mundo hay 350 millones de personas con infección crónica por hepatitis viral y 50 millones con cirrosis (7). En EE.UU., 4,5 millones de personas tienen HCV crónica y 29 millones han sido diagnosticadas de NAFLD. Hasta un tercio de las personas con cirrosis y entre 25-40 % con HBV desarrollarán
30 un HCC a lo largo de su vida, con un riesgo anual de hasta 8 % para los pacientes con cirrosis (8). Un grupo cada vez mayor de individuos con riesgo de cáncer hepático, incluidos 29

millones en EE.UU., tienen NAFLD y 20 % de los HCC que se desarrollan en esta población ocurren sin cirrosis (9). Las sociedades médicas de todo el mundo recomiendan la examinación para las poblaciones de mayor riesgo, actualmente con imágenes de ecografía abdominal con o sin alfafetoproteína (AFP, por sus siglas en inglés). Sin embargo, la adherencia general a las directrices internacionales sigue siendo baja: menos de 1 de cada 5 personas elegibles en todo el mundo recibe algún nivel de vigilancia y menos de 2 % sigue la examinación recomendada (10-12). Son muchos los factores que contribuyen a la escasa adherencia a las guías de examinación, incluyendo identificar los individuos de alto riesgo, el requisito de infraestructura y el personal necesario para los métodos de examinación basados en la imagenología (11). Las pruebas de examinación actuales que incluyen imagenología ecográfica, con o sin AFP, han mostrado una sensibilidad limitante que varía de 47-84 % con especificidades de 67 % a más de 90 % (13). Adicionalmente, la falta de enfoques diagnósticos no invasivos para la NAFLD sugiere que la población no cubierta actualmente por las recomendaciones de examinación del HCC está aumentando. Por lo tanto, existe una gran necesidad de desarrollar enfoques de examinación accesibles y sensibles para el HCC en todo el mundo.

Una vía reciente para superar estos retos ha sido el desarrollo de nuevos biomarcadores de ADN libre en de células (ADNlc) en sangre para la detección del cáncer. Los enfoques basados en mutaciones somáticas se han usado como biomarcadores del cáncer hepático, pero están limitados por la necesidad de identificar las mutaciones en los tejidos y por los pocos cambios detectables en el plasma (14). Los perfiles de metilación, tanto en sitios específicos como en el genoma completo, así como los cambios en el número de copias, también han proporcionado vías viables para la detección del cáncer hepático, pero su sensibilidad de detección en etapas muy tempranas de la enfermedad sigue siendo subóptima (15-20). Las pruebas de detección precoz de cánceres múltiples recientemente desarrolladas parecen útiles para la detección de muchos cánceres (incluido el cáncer hepático) en una cohorte de riesgo promedio (21), pero no hay informes publicados sobre el uso de estos métodos en una población de alto riesgo de HCC. Adicionalmente, el coste de la mayoría de las pruebas basadas en el ADNlc es mucho mayor a las estimaciones de lo que sería asequible para las pruebas de examinación en EE.UU. y en todo el mundo (22). La combinación de estos enfoques con AFP ha aumentado el rendimiento, pero requiere dos pruebas separadas y aún tiene limitaciones en la enfermedad en etapa temprana (23).

BREVE DESCRIPCIÓN DE LA INVENCION

En la presente se proporciona un análisis no invasivo y ultrasensible de moléculas individuales de ADN libre de células (ADNlc) para detectar cambios en los perfiles de fragmentación en el genoma. Se observó que los pacientes con cáncer hepático presentaban perfiles de fragmentación alterados como resultado de los cambios genómicos y cromatínicos del cáncer hepático, incluso de regiones asociadas a factores de transcripción específicos del hígado, en comparación con los sujetos sanos.

Por consiguiente, en un aspecto, un método para diagnosticar enfermedades o trastornos hepáticos en un sujeto, comprende aislar el ADN libre de células (ADNlc) circulante de un sujeto; realizar la secuenciación del genoma completo de moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación; comparar los perfiles de fragmentación con sujetos sanos y/o genoma de referencia, y diagnosticar si el sujeto tiene una enfermedad o trastorno hepático. En ciertas modalidades, los perfiles de fragmentación son consistentes para los sujetos sin cáncer mientras que los perfiles de fragmentación son muy variables para los sujetos con cáncer hepático. En ciertas modalidades, el método comprende además identificar los orígenes celulares de los perfiles de fragmentación del ADNlc, en donde identificar los orígenes celulares de los perfiles de fragmentación del ADNlc comprende comparar los perfiles del fragmentoma del genoma completo con la captura de la conformación cromosómica de secuenciación de alto rendimiento (Hi-C, por su nombre en inglés). En ciertas modalidades, los orígenes celulares de los perfiles de fragmentación del ADNlc de sujetos sanos se correlacionan con las células linfoblastoides como origen celular. En ciertas modalidades, los orígenes celulares de los perfiles de fragmentación del ADNlc de los sujetos con cáncer hepático se correlacionan con los perfiles de fragmentación del ADNlc de los compartimentos de cromatina de las células de sangre periférica. En ciertas modalidades, el método comprende además determinar si los perfiles de fragmentación del ADNlc se correlacionan con una alteración de la unión al ADN de los factores de transcripción. En ciertas modalidades, los sitios de unión al ADN del factor de transcripción se determinan al calcular la cobertura agregada del ADNlc en todos los sitios de unión al ADN del factor de transcripción identificados en comparación con la cobertura genómica adyacente global para producir una única métrica para cada factor de transcripción por muestra. En ciertas modalidades, los sitios de unión al ADN de factores de transcripción identificados en sujetos con cáncer hepático se comparan con los sitios de unión al ADN de factores de transcripción en sujetos sanos. En ciertas modalidades, los perfiles de fragmentación del ADNlc de los sujetos con cáncer hepático se correlacionan con una alteración de la unión al ADN de los factores de transcripción. En ciertas modalidades, el método comprende además determinar ganancias o pérdidas cromosómicas en sujetos con cáncer hepático en comparación con

sujetos sanos. En ciertas modalidades, el método comprende además un modelo de aprendizaje automático para determinar los cambios en los perfiles de fragmentación del ADNlc que clasifica al sujeto como paciente de cáncer basado en el perfil de fragmentación del ADNlc para el sujeto. En ciertas modalidades, el modelo de aprendizaje automático genera una puntuación para cada sujeto basado en una combinación de perfiles de fragmentación regional y a gran escala. En ciertas modalidades, la puntuación generada es el diagnóstico de las etapas del cáncer hepático.

En otro aspecto, se divulgan sistemas y métodos para detectar el cáncer al realizar la secuenciación del genoma completo de moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación y los datos introducidos en la memoria de la computadora; ejecutar un modelo de aprendizaje automático para determinar los cambios en los perfiles de fragmentación del ADNlc que clasifica al sujeto como paciente de cáncer basado en el perfil de fragmentación del ADNlc para el sujeto.

En otro aspecto, un método para diagnosticar el cáncer hepático, comprende aislar ADN libre de células (ADNlc) circulante de una muestra biológica y realizar la secuenciación del genoma completo de moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación; identificar los orígenes celulares de los perfiles de fragmentación del ADNlc que comprenden comparar los perfiles de fragmentación del genoma completo con la captura de conformación cromosómica de secuenciación de alto rendimiento (Hi-C); correlacionar los perfiles de fragmentación del ADNlc con la alteración de la unión al ADN de los factores de transcripción; determinar las ganancias o pérdidas cromosómicas en los sujetos con cáncer hepático en comparación con los sujetos sanos; ejecutar un modelo de aprendizaje automático para determinar los cambios en los perfiles de fragmentación del ADNlc que clasifica al sujeto como paciente con cáncer basado en el perfil de fragmentación del ADNlc para el sujeto; por lo mismo, diagnosticar el cáncer hepático y administrar un tratamiento contra el cáncer al sujeto. En ciertas modalidades, los orígenes celulares de los perfiles de fragmentación del ADNlc de sujetos sanos se correlacionan con las células linfoblastoides como origen celular. En ciertas modalidades, los orígenes celulares de los perfiles de fragmentación del ADNlc de los sujetos con cáncer hepático se correlacionan con los perfiles de fragmentación del ADNlc de los compartimentos de cromatina de las células de sangre periférica. En ciertas modalidades, los sitios de unión al ADN del factor de transcripción se determinan al calcular la cobertura agregada del ADNlc en todos los sitios de unión al ADN del factor de transcripción identificados en comparación con la cobertura genómica adyacente global para producir una única métrica para cada factor de transcripción por muestra. En

ciertas modalidades, los sitios de unión al ADN de factores de transcripción identificados en sujetos con cáncer hepático se comparan con los sitios de unión al ADN de factores de transcripción en sujetos sanos. En ciertas modalidades, los perfiles de fragmentación del ADNlc de los sujetos con cáncer hepático se correlacionan con una alteración de la unión al ADN de los factores de transcripción. En ciertas modalidades, el modelo de aprendizaje automático genera una puntuación para cada sujeto basado en una combinación de perfiles de fragmentación regional y a gran escala. En ciertas modalidades, la puntuación generada es el diagnóstico de las etapas del cáncer hepático. En ciertas modalidades, la puntuación generada es el diagnóstico de carcinoma hepatocelular (HCC). En ciertas modalidades, la puntuación generada es diferencialmente el diagnóstico de enfermedad hepática que comprende cirrosis. En ciertas modalidades, el tratamiento comprende: cirugía, quimioterapia adyuvante, quimioterapia neoadyuvante, radioterapia, terapia hormonal, terapia citotóxica, inmunoterapia, terapia de células T adoptiva, terapia dirigida y combinaciones de las mismas.

Definiciones

A menos que se defina lo contrario, todos los términos (incluidos los términos técnicos y científicos) usados en la presente tienen el mismo significado comúnmente entendido por un experto en la materia a la que pertenece esta invención. Se entenderá además que los términos, tales como los definidos en diccionarios de uso común, deben interpretarse en un sentido que sea coherente con su significado en el contexto de la materia pertinente y no se interpretarán en un sentido idealizado o excesivamente formal, a menos que así se defina expresamente en la presente.

Como se usa en la presente, las formas singulares “un”, “una” y “el/la” están destinadas también a incluir las formas plurales, a menos que el contexto indique claramente lo contrario. Además, en la medida en que los términos “incluyendo/que incluye”, “incluye”, “teniendo/que tiene”, “tiene”, “con”, o variantes de los mismos se usen en la descripción detallada y/o en las reivindicaciones, dichos términos están destinados a ser inclusivos de una manera similar al término “que comprende”.

El término “alrededor de” o “aproximadamente” significa dentro de un rango de error aceptable para el valor particular determinado por un experto en la materia, que dependerá en parte de cómo se mide o determina el valor, es decir, de las limitaciones del sistema de medición. Por ejemplo, “aproximadamente” puede significar dentro de 1 o más de 1 desviación estándar, de acuerdo con la práctica en la materia. Alternativamente, “aproximadamente” puede significar un rango de hasta 20 %, hasta 10 %, hasta 5 % o hasta 1 % de un valor o

rango determinado. Alternativamente, particularmente con respecto a los sistemas o procesos biológicos, el término puede significar dentro de un orden de magnitud dentro de 5 veces, y también dentro de 2 veces, de un valor. Cuando se describan valores particulares en la solicitud y en las reivindicaciones, a menos que se indique lo contrario, deberá suponerse que el término “aproximadamente” significa dentro de un rango de error aceptable para el valor particular.

Los términos “alineado”, “alineamiento”, “mapeado” o “alineación”, “mapeo” se refieren a una o más secuencias que se identifican como coincidentes en cuanto al orden de sus moléculas de ácido nucleico con una secuencia conocida de un genoma de referencia. Dicho alineamiento puede realizarse manualmente o por un algoritmo informático, los ejemplos incluyen el programa informático Efficient Local Alignment of Nucleotide Data (Alineamiento local eficiente de datos de nucleótidos, ELAND, por sus siglas en inglés) distribuido como parte de la canalización de Illumina Genomics Analysts. La coincidencia de una secuencia leída en el alineamiento puede ser de 100 % o menor a 100 % (coincidencia no perfecta).

El término “cáncer”, como se usa en la presente, se refiere a una enfermedad, condición, rasgo, genotipo o fenotipo caracterizado por un crecimiento o replicación celular no regulado, como se conoce en la materia. Los términos “neoplasia” y “tumor” se refieren a un tejido anormal que crece por proliferación celular más rápidamente de lo normal y que continúa creciendo después de eliminar el estímulo que inició la proliferación. Este tejido anormal muestra una falta parcial o total de organización estructural y de coordinación funcional con el tejido normal, que puede ser benigno (tal como un tumor benigno) o maligno (tal como un tumor maligno). Los ejemplos de cáncer incluyen cáncer hepático (incluido carcinoma hepatocelular [HCC]), I cáncer de pulmón (incluido carcinoma no microcítico de pulmón), cáncer gástrico, cáncer colorrectal, así como, por ejemplo, leucemias, por ejemplo, leucemia mielógena aguda (AML, por sus siglas en inglés), leucemia mielógena crónica (CML, por sus siglas en inglés), leucemia linfocítica aguda (ALL, por sus siglas en inglés) y leucemia linfocítica crónica, cánceres relacionados con el SIDA, tales como sarcoma de Kaposi; cánceres de mama cánceres óseos tales como osteosarcoma, condrosarcomas, sarcoma de Ewing, fibrosarcomas, tumores de células gigantes, adamantinomas y cordomas; cánceres cerebrales tales como meningiomas, glioblastomas, astrocitomas de bajo grado, oligodendrocitomas, tumores hipofisarios, schwannomas y cánceres cerebrales metastásicos; cánceres de cabeza y cuello, incluidos diversos linfomas tales como linfoma de células del manto, linfoma no Hodgkin, adenoma, carcinoma de células escamosas, carcinoma laríngeo, cánceres de vesícula biliar y vías biliares, cánceres de retina tales como retinoblastoma,

cánceres de esófago, cánceres gástricos, mieloma múltiple, cáncer de ovario, cáncer de útero, cáncer de tiroides, cáncer de testículo, cáncer de endometrio, melanoma, cáncer de vejiga, cáncer de próstata, cáncer de páncreas, sarcomas, tumor de Wilms, cáncer de cuello uterino, cáncer de cabeza y cuello, cánceres de piel, carcinoma nasofaríngeo, liposarcoma, carcinoma
5 epitelial, carcinoma de células renales, adenocarcinoma de vesícula biliar, adenocarcinoma de parótida, sarcoma de endometrio, cánceres multirresistentes; y enfermedades y condiciones proliferativas, tales como neovascularización asociada a la angiogénesis tumoral.

El término “ácido nucleico libre de células”, “polinucleótido libre de células”, “ADN libre de células” o “ADNlc” se refiere a fragmentos de ácido nucleico que circulan en el cuerpo de
10 un individuo (por ejemplo, en el torrente sanguíneo) y proceden de una o más células sanas y/o de una o más células cancerosas. Adicionalmente, el ADNlc puede proceder de otras fuentes tales como virus, fetos, etc.

El término “ADN tumoral circulante” o “ADNtc” se refiere a fragmentos de ácido nucleico procedentes de células tumorales u otros tipos de células cancerosas, que pueden
15 liberarse en el torrente sanguíneo de un individuo como resultado de procesos biológicos tales como la apoptosis o necrosis de células moribundas o liberados activamente por células tumorales viables.

Como se usa en la presente, los términos “que comprende”, “comprende” o “comprender”, y sus variaciones, en referencia a los elementos definidos o descritos de un
20 artículo, composición, aparato, método, proceso, sistema, etc., tienen la intención de ser inclusivos o abiertos, permitiendo elementos adicionales, indicando por lo mismo que el artículo, composición, aparato, método, proceso, sistema, etc., definido o descrito incluye esos elementos especificados, o, según proceda, equivalentes de los mismos, y que pueden incluirse otros elementos y aun así quedar dentro del alcance/definición del artículo,
25 composición, aparato, método, proceso, sistema, etc., definido.

“Diagnóstico” o “diagnosticado” significa identificar la presencia o naturaleza de una condición patológica. Los métodos de diagnóstico difieren en su sensibilidad y especificidad. La “sensibilidad” de un ensayo de diagnóstico es el porcentaje de individuos enfermos que dan positivo (porcentaje de “verdaderos positivos”). Los individuos enfermos no detectados
30 por el ensayo son “falsos negativos” Los sujetos que no están enfermos y que dan negativo en la prueba, se denominan “verdaderos negativos” La “especificidad” de un ensayo de diagnóstico es 1 menos la tasa de falsos positivos, en donde la tasa de “falsos positivos” se define como la proporción de personas sin la enfermedad que dan positivo en la prueba.

Aunque un método de diagnóstico particular puede no proporcionar un diagnóstico definitivo de una condición, basta con que el método proporcione una indicación positiva que ayude al diagnóstico.

Una “cantidad efectiva”, como se usa en la presente, significa una cantidad que
5 proporciona un beneficio terapéutico o profiláctico.

Como se usa en la presente, los términos “perfil de fragmentación”, “diferencias en los patrones de fragmentación en función de la posición” y “diferencias en el tamaño y la cobertura de los fragmentos en función de la posición a lo largo del genoma” son equivalentes y pueden usarse indistintamente. En algunas modalidades, determinar un perfil de fragmentación del
10 ADNlc en un mamífero puede usarse para identificar a un mamífero que tiene cáncer. Por ejemplo, los fragmentos de ADNlc obtenidos de un mamífero (por ejemplo, de una muestra obtenida de un mamífero) pueden someterse a una secuenciación del genoma completo de baja cobertura, y los fragmentos secuenciados pueden mapearse en el genoma (por ejemplo, en ventanas no solapadas) y valorarse para determinar un perfil de fragmentación del ADNlc.
15 Como se describe en la presente, el perfil de fragmentación del ADNlc de un mamífero con cáncer es más heterogéneo (por ejemplo, en la longitud de los fragmentos) que el perfil de fragmentación del ADNlc de un mamífero sano (por ejemplo, un mamífero sin cáncer). Como tal, esta divulgación también proporciona métodos y materiales para la valoración, monitorización y/o tratamiento de mamíferos (por ejemplo, humanos) que tienen, o se
20 sospecha que tienen, cáncer. En algunas modalidades, este documento proporciona métodos y materiales para identificar a un mamífero que tiene cáncer. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar la presencia y, opcionalmente, el tejido de origen del cáncer en el mamífero basándose, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas
25 modalidades, se proporcionan métodos y materiales para monitorizar si un mamífero tiene cáncer. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar la presencia del cáncer en el mamífero basándose, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas modalidades, se proporcionan métodos y materiales para identificar a un mamífero que tiene
30 cáncer y administrar uno o más tratamientos contra el cáncer al mamífero para tratar al mamífero. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede ser valorada para determinar si el mamífero tiene cáncer basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero, y uno o más tratamientos contra el cáncer pueden ser administrados al mamífero.

Como se usa en la presente, la “frecuencia” de mutaciones se define como el número de variantes por millón de posiciones evaluadas en todas las moléculas de ADN secuenciadas.

5 El término “ácido nucleico genómico” o “ADN genómico” se refiere al ácido nucleico que incluye el ADN cromosómico que se origina a partir de una o más células sanas (por ejemplo, no tumorales). En diversas modalidades, el ADN genómico puede extraerse de una célula derivada de un linaje celular sanguíneo, tal como un leucocito (WBC, por sus siglas en inglés).

10 Como se usa en la presente, el término “perfil mutacional” se refiere al tipo de mutación y a la frecuencia observada en paquetes a lo largo del genoma. Comparar los perfiles de mutación entre las regiones genómicas más comúnmente alteradas en el cáncer y los perfiles de mutación de las regiones más frecuentemente mutadas en el ADNlc normal puede usarse para determinar las diferencias multirregionales.

15 “Opcional” u “opcionalmente” significa que el evento o circunstancia descrito subsecuentemente puede o no ocurrir, y que la descripción incluye casos en los que el evento o circunstancia ocurre y casos en los que no.

Como se usa en esta memoria descriptiva y en las reivindicaciones anexas, el término “o” se emplea generalmente en su sentido que incluye “y/o” a menos que el contenido dicte claramente lo contrario.

20 La administración “parenteral” de una composición inmunogénica incluye, por ejemplo, la inyección subcutánea (s.c.), intravenosa (i.v.), intramuscular (i.m.) o intraesternal, o técnicas de infusión.

25 Los términos “paciente” o “individuo” o “sujeto” se usan indistintamente en la presente, y se refieren a un sujeto mamífero a tratar, siendo preferibles los pacientes humanos. En algunas modalidades, los métodos de la invención se usan en animales de experimentación, en aplicaciones veterinarias y en el desarrollo de modelos animales de enfermedades, incluyendo, pero no se limitan a, roedores como ratones, ratas, hámsters y primates.

30 El término “genoma de referencia”, como se usa en la presente, puede referirse a una base de datos de secuencias de ácidos nucleicos digital o previamente identificada, ensamblada como ejemplo representativo de una especie o sujeto. Los genomas de referencia pueden ensamblarse a partir de las secuencias de ácido nucleico de múltiples

sujetos, muestras u organismos y no representan necesariamente la composición de ácido nucleico de una única persona. Los genomas de referencia pueden usarse para mapear las lecturas de secuenciación de una muestra en posiciones cromosómicas. Por ejemplo, en el National Center for Biotechnology Information (Centro Nacional de Información Biotecnológica, NCBI, por sus siglas en inglés), en ncbi.nlm.nih.gov, se encuentra un genoma de referencia usado para sujetos humanos, así como para muchos otros organismos.

El término “segmento de lectura” o “lectura” se refiere a cualquier secuencia de nucleótidos, incluidas las lecturas de secuencias obtenidas de un individuo y/o las secuencias de nucleótidos derivadas de la secuencia inicial leída de una muestra obtenida de un individuo.

Los términos “muestra”, “muestra de paciente”, “muestra biológica” y similares abarcan una variedad de tipos de muestras obtenidas de un paciente, individuo o sujeto y que pueden usarse en un ensayo de diagnóstico, pronóstico y/o monitorización. La muestra del paciente puede obtenerse de un sujeto sano, de un paciente enfermo o de un paciente con cáncer de pulmón. En ciertas modalidades, una muestra “proporcionada” puede ser obtenida por la persona (o máquina) que realiza el ensayo, o puede haber sido obtenida por otra, y transferida a la persona (o máquina) que realiza el ensayo. Asimismo, una muestra obtenida de un paciente puede dividirse y usar solamente una parte para el diagnóstico. Además, la muestra, o una porción de la misma, puede almacenarse en condiciones que permitan mantener la muestra para su posterior análisis. La definición abarca específicamente la sangre y otras muestras líquidas de origen biológico (incluyendo, pero no se limita a, sangre periférica, suero, plasma, sangre del cordón umbilical, líquido amniótico, líquido cefalorraquídeo, orina, saliva, heces y líquido sinovial), muestras de tejidos sólidos tales como una muestra de biopsia o cultivos de tejidos o células derivadas de los mismos y su progenie. En cierta modalidad, una muestra comprende líquido cefalorraquídeo. En una modalidad específica, una muestra comprende una muestra de sangre. En otra modalidad, una muestra comprende una muestra de plasma. En otra modalidad, se usa una muestra de suero. La definición de “muestra” también incluye las muestras que han sido manipuladas de cualquier forma después de su obtención, tal como por centrifugación, filtración, precipitación, diálisis, cromatografía, tratamiento con reactivos, lavado o enriquecimiento para ciertas poblaciones celulares. Los términos abarcan además una muestra clínica, y también incluyen células en cultivo, sobrenadantes celulares, muestras de tejidos, órganos y similares. Las muestras también pueden comprender bloques de tejido recién congelados y/o fijados con formalina e incluidos en parafina, tales como bloques preparados a partir de biopsias clínicas o patológicas, preparados para el análisis patológico o el estudio mediante inmunohistoquímica.

El término “lecturas de secuencias” se refiere a las secuencias de nucleótidos leídas a partir de una muestra obtenida de un individuo. Las lecturas de secuencias pueden obtenerse a través de diversos métodos conocidos en la materia.

5 Como se define en la presente, una cantidad “terapéuticamente efectiva” de un compuesto o agente (es decir, una dosificación efectiva) significa una cantidad suficiente para producir un resultado terapéuticamente (por ejemplo, clínicamente) deseable. Las composiciones pueden administrarse desde una o más veces al día hasta una o más veces a la semana, incluso una vez cada dos días. El experto apreciará que ciertos factores pueden influir en la dosificación y el tiempo necesarios para tratar efectivamente a un sujeto,
10 incluyendo pero no limitado a la gravedad de la enfermedad o trastorno, los tratamientos previos, la salud general y/o la edad del sujeto, y otras enfermedades presentes. Asimismo, el tratamiento de un sujeto con una cantidad terapéuticamente efectiva de los compuestos de la invención puede incluir un único tratamiento o una serie de tratamientos.

Como se usa en la presente, los términos “tratar”, “tratamiento” y similares se refieren
15 a reducir o mejorar un trastorno y/o los síntomas asociados con el mismo. Se apreciará que, aunque no se excluye, el tratamiento de un trastorno o condición no requiere que el trastorno, condición o síntomas asociados con el mismo sean completamente eliminados.

Genes: Todos los genes, nombres de genes y productos génicos divulgados en la presente están contemplados para corresponder a homólogos de cualquier especie a la que
20 sean aplicables las composiciones y métodos divulgados en la presente. Se entiende que cuando se divulga un gen o un subproducto génico de una especie particular, esta divulgación solamente se contempla a título de ejemplo y no debe interpretarse como una limitación, a menos que el contexto en el que aparece lo indique claramente. Así, por ejemplo, en el caso de los genes o productos génicos divulgados en la presente, se pretende abarcar genes y
25 productos génicos homólogos y/u ortólogos de otras especies.

Rangos: a lo largo de esta divulgación, diversos aspectos de la invención pueden presentarse en un formato de rango. Debe entenderse que la descripción en formato de rango es meramente por conveniencia y brevedad y no debe interpretarse como una limitación inflexible del alcance de la invención. En consecuencia, debe considerarse que la descripción
30 de un rango divulga específicamente todos los posibles subrangos, así como los valores numéricos individuales dentro de ese rango. Por ejemplo, debe considerarse que la descripción de un rango tal como de 1 a 6 ha divulgado específicamente subrangos tales como de 1 a 3, de 1 a 4, de 1 a 5, de 2 a 4, de 2 a 6, de 3 a 6, etc., así como números

individuales dentro de ese rango, por ejemplo, 1, 2, 2,7, 3, 4, 5, 5,3 y 6. Esto aplica independientemente de la amplitud del rango.

Cualquiera de las composiciones o métodos proporcionados en la presente pueden combinarse con uno o más de cualquiera de las otras composiciones y métodos proporcionados en la presente.

BREVE DESCRIPCIÓN DE LAS FIGURAS

El archivo de la patente o de la solicitud contiene al menos un dibujo realizado en color. La Oficina facilitará copias de esta patente o publicación de solicitud de patente con dibujos en color previa solicitud y pago de la tasa correspondiente.

La **figura 1** (incluye las **figuras 1A-1C**) son gráficas que demuestran que los perfiles de fragmentación del genoma completo reflejan la estructura subyacente de la cromatina. **Figura 1A:** Perfiles de fragmentación de 501 individuos en 473 regiones genómicas de 5 mb no solapadas. Los perfiles de fragmentación de los individuos con cáncer muestran una marcada heterogeneidad en comparación con los individuos sin cáncer con y sin enfermedad hepática. **Figura 1B:** Comparación de los atributos de fragmentación del plasma con los compartimentos A/B de referencia. El canal 1 muestra compartimentos A/B extraídos de tejido de cáncer hepático (29). El canal 2 muestra la mediana de componente de cáncer hepático extraído de las muestras de plasma de HCC de diez pacientes hepáticos con alta fracción tumoral por ichorCAN (por sus siglas en inglés) (57). El canal 3 muestra la mediana del perfil de fragmentación en el plasma para estas 10 muestras de HCC y el canal 4 muestra la mediana del perfil para 10 muestras de plasma sano. El canal 5 muestra los compartimentos A/B de las células linfoblásticas (29). Estos cinco pistas muestran el cromosoma 22 como ejemplo, con un sombreado más oscuro que indica las regiones informativas del genoma en las que las dos pistas de referencia difieren en dominio (abierto/cerrado) o magnitud. **Figura 1C:** muestra otros resultados de comparar los atributos de fragmentación del plasma con los compartimentos A/B de referencia.

La **figura 2** (incluye las **figuras 2A-2E**) son una serie de gráficas que muestran los perfiles de fragmentación en pacientes con HCC que destacan los factores de transcripción específicos del hígado. **Figura 2A:** La cobertura en y alrededor de los sitios de unión a factores de transcripción (TFBS, por sus siglas en inglés) para los 9 TF (por sus siglas en inglés) para los que la cobertura relativa en el sitio de unión que tenía la mayor separación de HCC de muestras sin cáncer. Se representa la media de cada grupo, con +/- una desviación estándar

(SD, por sus siglas en inglés) sombreada. Estos intervalos de confianza (CI, por sus siglas en inglés) muestran una separación, lo que pone de relieve que las diferencias de cobertura en un TFBS pueden proporcionar información sobre el estado del cáncer. **Figura 2B:** La cobertura en y alrededor del TFBS para los 9 TF que tuvieron la menor separación de HCC de muestras no cancerosas en la cohorte EE.UU./UE. Estos CI se solapan en gran medida, lo que refleja su condición de TFBS con escasa discriminación. El análisis de enriquecimiento de conjuntos de genes de los TF analizados tanto en el carcinoma hepatocelular como en el adenocarcinoma de pulmón mostró que los TF están selectivamente enriquecidos en numerosas rutas relacionadas con el cáncer hepático y de pulmón respectivamente (**figura 2C**), incluidos el carcinoma hepático adulto y el adenocarcinoma de pulmón (**figura 2D, figura 2E**).

La **figura 3** (incluye las **figuras 3A-3C**) muestra que los atributos de fragmentación de alta dimensión reflejan la biología del cáncer hepático y se incorporan en los enfoques de aprendizaje automático DELFI (por sus siglas en inglés). **Figura 3A:** Un mapa de calor que refleja la complejidad de la fragmentación del genoma completo y los atributos del sitio de unión del factor de transcripción usados en el enfoque de aprendizaje automático DELFI. Cada fila representa una muestra, mientras que las columnas muestran atributos genómicos individuales. **Figura 3B:** El análisis de los cambios en el número de copias en tejido de 372 cánceres hepáticos TCGA (por sus siglas en inglés) y plasma de 501 individuos refleja la consistencia biológica. Los cambios en el número de copias que se producen en TCGA (rojo = ganancias, azul = pérdidas) también se encontraron a nivel de brazo cromosómico en el plasma de HCC, pero no en individuos sin cáncer. **Figura 3C:** Mapa de calor que representa las contribuciones de las regiones genómicas individuales al modelo DELFI final entrenado. Los atributos de fragmentación se resumieron como tres componentes principales en el modelo, mientras que la aneuploidía se resumió como puntuaciones z a nivel de brazo. Los paneles superior, central y derecho muestran los coeficientes de los componentes de fragmentación, las puntuaciones z a nivel de brazo y los sitios de unión de factores de transcripción, respectivamente, en el modelo. CNV, variación del número de copias (por sus siglas en inglés).

La **figura 4** (incluye las **figuras 4A-4D**) son una serie de gráficas que demuestran que los modelos de aprendizaje automático DELFI detectan el cáncer hepático con una alta sensibilidad y especificidad. **Figura 4A:** Puntuaciones DELFI para la cohorte EE.UU./UE en función de la enfermedad hepática y la etapa del cáncer para los modelos de examinación y vigilancia. Los pacientes cirróticos tienen puntuaciones DELFI mayores a las de los individuos

sin cáncer o con hepatitis viral en promedio, pero menores a las de todas las etapas del cáncer hepático. Los pacientes con cáncer hepático en todas las etapas presentan puntuaciones DELFI relativamente altas, siendo los individuos en etapa C los que presentan uniformemente las puntuaciones DELFI más altas. **Figura 4B:** Análisis ROC (por sus siglas en inglés) de la cohorte de población general de EE.UU./UE y de la cohorte de vigilancia de alto riesgo. **Figura 4C:** Análisis ROC de la población general de EE.UU./UE y cohortes de vigilancia separadas por la etapa BCLC (por sus siglas en inglés), mostrando una alta sensibilidad y especificidad en todas las etapas. **Figura 4D:** Análisis ROC para el modelo de vigilancia fijo aplicado a la cohorte Hong Kong, que incluye 90 individuos con HCC (85 con cáncer en etapa A de BCLC y 5 con cáncer en etapa B de BCLC), 101 individuos con cirrosis y hepatitis viral y 32 individuos sin cáncer ni enfermedad hepática.

La **figura 5** es un mapa de calor que representa las contribuciones de las regiones genómicas individuales al modelo de vigilancia DELFI final entrenado. Los atributos de fragmentación se resumieron como tres componentes principales en el modelo, mientras que la aneuploidía se resumió como puntuaciones z a nivel de brazo. Los paneles superior y derecho muestran la importancia variable de los componentes de fragmentación y las puntuaciones z a nivel de brazo, respectivamente, en el modelo.

La **figura 6A** y la **figura 6B** son gráficas que muestran que las puntuaciones DELFI en individuos sin cáncer no se ven afectados por la edad y el sexo. Puntuaciones DELFI en individuos (n = 293, 45 mujeres, 88 hombres) sin cáncer ni enfermedad hepática por edad (**figura 6A**) o el sexo (**figura 6B**).

La **figura 7** es una serie de gráficas que demuestran que las puntuaciones DELFI en individuos sin cáncer no son diferentes entre grupos raciales o étnicos.

La **figura 8** es un gráfica que demuestra que las puntuaciones DELFI en pacientes cirróticos (n = 40) con información disponible no estaban relacionados con las puntuaciones Child-Pugh de gravedad de la cirrosis.

La **figura 9** es una serie de gráficas que demuestran que las puntuaciones DELFI en individuos sin cáncer (n = 53 en pacientes con hepatitis viral, y n = 78 en pacientes cirróticos) varían con el IMC.

La **figura 10A** y la **figura 10B** muestran el rendimiento de los modelos DELFI alternativos con y sin la adición de los atributos de cobertura relativa del sitio de unión del TF.

Análisis ROC del modelo de vigilancia que incorpora atributos de sitios de unión de factores de transcripción procedentes de análisis CHIP-Seq derivados del hígado en la cohorte de alto riesgo de EE.UU./UE, y para el modelo de examinación sin sitios de unión de factores de transcripción para la cohorte de población general de EE.UU./UE (**figura 10A**), y por etapa de BCLC (**figura 10B**).

La **figura 11** muestra un gráfica que demuestra la correlación entre las puntuaciones DELFI ordenadas por rango de los modelos de Vigilancia y Examinación. Existe una alta correlación entre las puntuaciones ordenadas por rango usando nuestros modelos DELFI de alto riesgo y de examinación entre todos los pacientes con HCC (n = 75).

La **figura 12A** y la **figura 12B** son una serie de gráficas que demuestran que las puntuaciones DELFI en pacientes con HCC estaban relacionados con el tamaño y el número de la lesión tumoral. Las puntuaciones DELFI de los pacientes con HCC (n = 75) se correlacionaron positivamente con el aumento del diámetro de la lesión (**figura 12A**) y mostraron una tendencia positiva al aumentar el número de lesiones (**figura 12B**).

La **figura 13** es un gráfica que demuestra que las puntuaciones DELFI en pacientes con HCC (n = 54) con enfermedad en etapa resecable (BCLC 0, A y B) no mostraron diferencias por la etiología de la enfermedad hepática subyacente. Los pacientes en etapa C (n = 21) no se incluyeron en este análisis, ya que estos pacientes tenían una etiología relacionada en gran medida con la hepatitis viral (n = 14).

La **figura 14** es un gráfica que demuestra que la puntuación DELFI en pacientes con HCC está correlacionado con los niveles de AFP. En total, 39 de los 75 individuos con HCC tenían niveles de AFP mayores al umbral de examinación recomendado de 20 ng/mL, indicado por la línea de puntos vertical. Todos menos tres de estos individuos habrían sido detectados en el umbral DELFI correspondiente a una especificidad de 80 % en la cohorte de alto riesgo (0,26), indicada por la línea punteada horizontal. Entre los individuos que no habrían sido detectados por AFP, DELFI detectó 30 de 36.

La **figura 15** es un gráfica que demuestra que la correlación de los perfiles de fragmentación con la mediana de los perfiles no cancerosos no difiere entre los subgrupos no cancerosos.

La **figura 16** es un gráfica que demuestra que los perfiles de fragmentación del genoma completo en la cohorte de validación de Hong Kong muestran similitudes con los de

la cohorte EE.UU./UE. Perfiles de fragmentación de 223 individuos en 473 regiones genómicas de 5 mb no solapadas. Los perfiles de fragmentación de los individuos con cáncer muestran una marcada heterogeneidad en comparación con los individuos sin cáncer con y sin enfermedad hepática, similar a lo observado en la cohorte EE.UU./UE.

5 La **figura 17** muestra que los cambios en el número de copias cromosómicas detectados en plasma son biológicamente consistentes en las cohortes de EE.UU./UE y Hong Kong, así como en las muestras de tejido TCGA. El rojo indica ganancias en el número de copias, mientras que el azul indica pérdidas en el número de copias. Estos cambios se observan fácilmente tanto en el tejido como en el plasma del HCC, pero no son evidentes en
10 individuos sin cáncer. CNV, variación del número de copias.

La **figura 18** (incluye las **figuras 18A-18D**) muestra la modelización de la aplicación del DELFI en la examinación del cáncer hepático. **Figura 18A:** Se modeló la incertidumbre de la sensibilidad y especificidad de la ecografía y la AFP, así como de la examinación DELFI, en una población teórica de 100.000 individuos de alto riesgo. Distribuciones predictivas para
15 el número de cánceres hepáticos detectados (**figura 18B**), tasa de falsos negativos (**figura 18C**), y valores negativos predictivos (**figura 18D**) entre estos individuos incorporó la variación tanto en la prevalencia del cáncer hepático como en la adherencia a la examinación basada en imágenes y sangre. La línea central de los gráficos de caja representa la mediana, el límite superior de los gráficos de caja representa el tercer cuantil (percentil 75), el límite inferior de
20 los gráficos de caja representa el primer cuantil (percentil 25), el borde superior es el valor máximo de los datos que se encuentra dentro de 1,5 veces el rango intercuartílico por encima del percentil 75, y el borde inferior es el valor mínimo de los datos que se encuentra dentro de 1,5 veces el rango intercuartílico por debajo del percentil 25. Muestra que los cambios en el número de copias cromosómicas detectados en plasma son biológicamente consistentes en
25 las cohortes de EE.UU./UE y Hong Kong, así como en las muestras de tejido TCGA. El rojo indica ganancias en el número de copias, mientras que el azul indica pérdidas en el número de copias. Estos cambios se observan fácilmente tanto en el tejido como en el plasma del HCC, pero no son evidentes en individuos sin cáncer. CNV, variación del número de copias.

DESCRIPCIÓN DETALLADA DE LA INVENCION

30 DELFI (DNA evaluation of fragments for early interception, Evaluación de fragmentos de ADN para su interceptación precoz) usa perfiles de fragmentación del genoma completo para proporcionar un enfoque de alto rendimiento y rentable para la detección del cáncer (21, 22). La información sobre fragmentación y metilación también ha demostrado su capacidad para

diferenciar a los pacientes con cáncer hepático de los que no lo padecen (24), aunque tal enfoque requiere dos métodos distintos de preparación y análisis de bibliotecas de ADNlc. Hasta la fecha, ningún estudio ha validado los enfoques genómicos para detectar el HCC en grupos independientes o en diferentes poblaciones de alto riesgo.

5 Por consiguiente, en ciertas modalidades, un método para diagnosticar enfermedades o trastornos hepáticos en un sujeto comprende aislar ADN libre de células (ADNlc) circulante de un sujeto; realizar la secuenciación del genoma completo de moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación; comparar los perfiles de fragmentación con sujetos sanos y/o genoma de referencia, y diagnosticar si el sujeto tiene
10 una enfermedad o trastorno hepático.

 En otra modalidad, un método para diagnosticar el cáncer hepático comprende aislar el ADN libre de células (ADNlc) circulante de una muestra biológica y realizar la secuenciación del genoma completo de las moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación; identificar los orígenes celulares de los perfiles de fragmentación
15 del ADNlc que comprende comparar los perfiles de fragmentación del genoma completo con la captura de conformación cromosómica de secuenciación de alto rendimiento (Hi-C); correlacionar los perfiles de fragmentación del ADNlc con la alteración de la unión al ADN de los factores de transcripción; determinar las ganancias o pérdidas cromosómicas en los sujetos con cáncer hepático en comparación con los sujetos sanos; ejecutar un modelo de
20 aprendizaje automático para determinar los cambios en los perfiles de fragmentación del ADNlc que clasifica al sujeto como paciente con cáncer basado en el perfil de fragmentación del ADNlc para el sujeto; por lo mismo, diagnosticar el cáncer hepático y administrar un tratamiento contra el cáncer al sujeto.

Perfiles de fragmentación del ADNlc

25 Un perfil de fragmentación del ADNlc puede incluir uno o más patrones de fragmentación del ADNlc. Un patrón de fragmentación del ADNlc puede incluir cualquier patrón de fragmentación del ADNlc apropiado. Los ejemplos de patrones de fragmentación del ADNlc incluyen, sin limitación, la mediana del tamaño de los fragmentos, la distribución del tamaño de los fragmentos, la relación entre fragmentos pequeños de ADNlc y fragmentos
30 grandes de ADNlc, y la cobertura de los fragmentos de ADNlc. En algunas modalidades, un patrón de fragmentación del ADNlc incluye dos o más (por ejemplo, dos, tres o cuatro) de la mediana del tamaño de los fragmentos, la distribución del tamaño de los fragmentos, la relación entre fragmentos pequeños de ADNlc y fragmentos grandes de ADNlc, y la cobertura

de los fragmentos de ADNlc. En algunas modalidades, el perfil de fragmentación del ADNlc puede ser un perfil de ADNlc del genoma completo (por ejemplo, un perfil de ADNlc del genoma completo en ventanas a lo largo del genoma). En algunas modalidades, el perfil de fragmentación del ADNlc puede ser un perfil de región objetivo. Una región objetivo puede ser cualquier porción apropiada del genoma (por ejemplo, una región cromosómica). Los ejemplos de regiones cromosómicas para las que puede determinarse un perfil de fragmentación del ADNlc como se describe en la presente incluyen, sin limitación, una porción de un cromosoma (por ejemplo, una porción de 2q, 4p, 5p, 6q, 7p, 8q, 9q, 10q, 11q, 12q, y/o 14q) y un brazo cromosómico (por ejemplo, un brazo cromosómico de 8q,13q, 11q, y/o 3p). En algunas modalidades, un perfil de fragmentación del ADNlc puede incluir dos o más perfiles de regiones objetivo.

En algunas modalidades, se puede usar un perfil de fragmentación del ADNlc para identificar cambios (por ejemplo, alteraciones) en las longitudes de los fragmentos de ADNlc. Una alteración puede ser una alteración del genoma completo o una alteración en una o más regiones o *loci* objetivo. Una región objetivo puede ser cualquier región que contenga una o más alteraciones específicas del cáncer. En algunas modalidades, puede usarse un perfil de fragmentación del ADNlc para identificar (por ejemplo, identificar simultáneamente) de aproximadamente 10 alteraciones a aproximadamente 500 alteraciones (por ejemplo, de aproximadamente 25 a aproximadamente 500, de aproximadamente 50 a aproximadamente 500, de aproximadamente 100 a aproximadamente 500, de aproximadamente 200 a aproximadamente 500, de aproximadamente 300 a aproximadamente 500, de aproximadamente 10 a aproximadamente 400, de aproximadamente 10 a aproximadamente 300, de aproximadamente 10 a aproximadamente 200, de aproximadamente 10 a aproximadamente 100, de aproximadamente 10 a aproximadamente 50, de aproximadamente 20 a aproximadamente 400, de aproximadamente 30 a aproximadamente 300, de aproximadamente 40 a aproximadamente 200, de aproximadamente 50 a aproximadamente 100, de aproximadamente 20 a aproximadamente 100, de aproximadamente 25 a aproximadamente 75, de aproximadamente 50 a aproximadamente 250, o de aproximadamente 100 a aproximadamente 200, alteraciones).

El perfil de fragmentación del ADNlc puede obtenerse usando cualquier método apropiado. En algunas modalidades, el ADNlc de un mamífero (por ejemplo, un mamífero que tiene o se sospecha que tiene cáncer) puede procesarse en bibliotecas de secuenciación que pueden someterse a secuenciación del genoma completo (por ejemplo, secuenciación del genoma completo de baja cobertura), mapearse en el genoma y analizarse para determinar

las longitudes de los fragmentos de ADNlc. Las secuencias mapeadas pueden analizarse en ventanas no solapadas que cubren el genoma. Las ventanas pueden ser de cualquier tamaño. Por ejemplo, las ventanas pueden tener de miles a millones de bases. Como ejemplo no limitante, una ventana puede tener una longitud aproximada de 5 megabases (Mb). Se puede
5 asignar cualquier número apropiado de ventanas. Por ejemplo, se pueden mapear de decenas a miles de ventanas en el genoma. Por ejemplo, se pueden mapear cientos o miles de ventanas en el genoma. Dentro de cada ventana puede determinarse un perfil de fragmentación del ADNlc.

En algunas modalidades, los métodos y materiales descritos en la presente también
10 pueden incluir el aprendizaje automático. Por ejemplo, se puede usar el aprendizaje automático para identificar las frecuencias de mutación, el perfil de fragmentación alterado (por ejemplo, usando la cobertura de los fragmentos de ADNlc, el tamaño de los fragmentos de ADNlc, la cobertura de los cromosomas y el ADNmt).

En algunas modalidades, determinar un perfil de fragmentación del ADNlc en un
15 mamífero puede usarse para identificar a un mamífero que tiene cáncer. Por ejemplo, los fragmentos de ADNlc obtenidos de un mamífero (por ejemplo, de una muestra obtenida de un mamífero) pueden someterse a una secuenciación del genoma completo de baja cobertura, y los fragmentos secuenciados pueden mapearse en el genoma y valorarse para determinar un perfil de fragmentación del ADNlc. Como se describe en la presente, el perfil de fragmentación
20 del ADNlc de un mamífero con cáncer es más heterogéneo (por ejemplo, en la longitud de los fragmentos) que el perfil de fragmentación del ADNlc de un mamífero sano (por ejemplo, un mamífero sin cáncer). Como tal, también se proporcionan métodos y materiales para la valoración, monitorización y/o tratamiento de mamíferos (por ejemplo, humanos) que tienen, o se sospecha que tienen, cáncer. En algunas modalidades, se proporcionan métodos y
25 materiales para identificar a un mamífero que tiene cáncer. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar la presencia y, opcionalmente, el tejido de origen del cáncer en el mamífero basándose, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas modalidades, se proporcionan métodos y materiales para monitorizar si un mamífero tiene
30 cáncer. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar la presencia del cáncer en el mamífero basándose, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas modalidades, se proporcionan métodos y materiales para identificar a un mamífero que tiene cáncer y administrarle uno o más tratamientos contra el cáncer para tratar al mamífero. Por

ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede ser valorada para determinar si el mamífero tiene cáncer basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero, y uno o más tratamientos contra el cáncer pueden ser administrados al mamífero.

5 En algunas modalidades, se puede usar un perfil de fragmentación del ADNlc para detectar ADN derivado de tumores. Por ejemplo, se puede usar un perfil de fragmentación del ADNlc para detectar ADN derivado de tumores al comparar un perfil de fragmentación del ADNlc de un mamífero que tiene, o se sospecha que tiene, cáncer con un perfil de fragmentación del ADNlc de referencia (por ejemplo, un perfil de fragmentación del ADNlc de un mamífero sano y/o un perfil de fragmentación de ADN nucleosómico de células sanas del mamífero que tiene, o se sospecha que tiene, cáncer). En algunas modalidades, un perfil de fragmentación del ADNlc de referencia es un perfil generado previamente a partir de un mamífero sano. Por ejemplo, los métodos proporcionados en la presente pueden usarse para determinar un perfil de fragmentación del ADNlc de referencia en un mamífero sano, y ese perfil de fragmentación del ADNlc de referencia puede almacenarse (por ejemplo, en una computadora u otro medio de almacenamiento electrónico) para compararlo en el futuro con un perfil de fragmentación del ADNlc de prueba en un mamífero que tiene, o se sospeche que tiene, cáncer. En algunas modalidades, se determina un perfil de fragmentación del ADNlc de referencia (por ejemplo, un perfil de fragmentación del ADNlc almacenado) de un mamífero sano en el genoma completo. En algunas modalidades, se determina un perfil de fragmentación del ADNlc de referencia (por ejemplo, un perfil de fragmentación del ADNlc almacenado) de un mamífero sano en un intervalo subgenómico.

En algunas modalidades, un perfil de fragmentación del ADNlc puede usarse para identificar a un mamífero (por ejemplo, un humano) que tiene cáncer (por ejemplo, un cáncer hepático, un cáncer colorrectal, un cáncer de pulmón, un cáncer de mama, un cáncer gástrico, un cáncer de páncreas, un cáncer de vías biliares y/o un cáncer de ovario).

Un perfil de fragmentación del ADNlc puede incluir un patrón de tamaño de fragmentos de ADNlc. Los fragmentos de ADNlc pueden ser de cualquier tamaño apropiado. Por ejemplo, el fragmento de ADNlc puede tener una longitud de aproximadamente 50 pares de bases (pb) a aproximadamente 400 pb.

Un perfil de fragmentación del ADNlc puede incluir una distribución del tamaño de los fragmentos de ADNlc. Como se describe en la presente, un mamífero con cáncer puede tener una distribución de tamaños del ADNlc más variable que una distribución de tamaños de

fragmentos de ADNlc en un mamífero sano. En algunas modalidades, una distribución de tamaño puede estar dentro de una región objetivo. Un mamífero sano (por ejemplo, un mamífero que no tiene cáncer) puede tener una distribución de tamaños de fragmentos de ADNlc de la región objetivo de aproximadamente 1 o menor a aproximadamente 1. En algunas modalidades, un mamífero con cáncer puede tener una distribución de tamaños de fragmentos de ADNlc de la región objetivo más larga (por ejemplo, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más pb más larga, o cualquier número de pares de bases entre estos números) que una distribución de tamaños de fragmentos de ADNlc de la región objetivo en un mamífero sano. En algunas modalidades, un mamífero con cáncer puede tener una distribución de tamaños de fragmentos de ADNlc de la región objetivo más corta (por ejemplo, 10, 15, 20, 25, 30, 35, 40, 45, 50 o más pb más corta, o cualquier número de pares de bases entre estos números) que una distribución de tamaños de fragmentos de ADNlc de la región objetivo en un mamífero sano. En algunas modalidades, una distribución de tamaño puede ser una distribución de tamaño del genoma completo. Un mamífero sano (por ejemplo, un mamífero que no tiene cáncer) puede tener distribuciones muy similares de fragmentos de ADNlc cortos y largos en el genoma completo. En algunas modalidades, un mamífero con cáncer puede tener, en el genoma completo, una o más alteraciones (por ejemplo, aumentos y disminuciones) en los tamaños de los fragmentos de ADNlc. La una o más alteraciones pueden ser cualquier región cromosómica apropiada del genoma. Por ejemplo, una alteración puede producirse en una porción de un cromosoma. Los ejemplos de porciones de cromosomas que pueden contener una o más alteraciones en los tamaños de los fragmentos de ADNlc incluyen, sin limitación, porciones de 2q, 4p, 5p, 6q, 7p, 8q, 9q, 10q, 11q, 12q y 14q. Por ejemplo, una alteración puede producirse en un brazo cromosómico (por ejemplo, en todo un brazo cromosómico).

Un perfil de fragmentación del ADNlc puede incluir una relación entre fragmentos pequeños de ADNlc y fragmentos grandes de ADNlc y una correlación de las relaciones de fragmentos con las relaciones de fragmentos de referencia. Como se usa en la presente, con respecto a las relaciones entre fragmentos pequeños de ADNlc y fragmentos grandes de ADNlc, un fragmento pequeño de ADNlc puede tener una longitud de aproximadamente 100 pb a aproximadamente 150 pb. Como se usa en la presente, con respecto a las relaciones entre fragmentos pequeños de ADNlc y fragmentos grandes de ADNlc, un fragmento grande de ADNlc puede tener una longitud de aproximadamente 151 pb a 220 pb. Un mamífero con cáncer puede tener una correlación de relaciones de fragmentos (por ejemplo, una correlación de relaciones de fragmentos de ADNlc con respecto a relaciones de fragmentos de ADN de referencia, tales como relaciones de fragmentos de ADN de uno o más mamíferos sanos) que

es menor (por ejemplo, 2 veces menor, 3 veces menor, 4 veces menor, 5 veces menor, 6 veces menor, 7 veces menor, 8 veces menor, 9 veces menor, 10 veces menor, o más) que en un mamífero sano. Un mamífero sano (por ejemplo, un mamífero que no tiene cáncer) puede tener una correlación de relaciones de fragmentos (por ejemplo, una correlación de las relaciones de fragmentos de ADNlc con las relaciones de fragmentos de ADN de referencia, tales como las relaciones de fragmentos de ADN de uno o más mamíferos sanos) de aproximadamente 1 (por ejemplo, aproximadamente 0,96). En algunas modalidades, un mamífero con cáncer puede tener una correlación de relaciones de fragmentos (por ejemplo, una correlación de relaciones de fragmentos de ADNlc con respecto a relaciones de fragmentos de ADN de referencia, tales como relaciones de fragmentos de ADN de uno o más mamíferos sanos) que es, en promedio, menor a una correlación de relaciones de fragmentos (por ejemplo, una correlación de relaciones de fragmentos de ADNlc con respecto a relaciones de fragmentos de ADN de referencia, tales como relaciones de fragmentos de ADN de uno o más mamíferos sanos) en un mamífero sano.

Un perfil de fragmentación del ADNlc puede incluir la cobertura de todos los fragmentos. La cobertura de todos los fragmentos puede incluir ventanas (por ejemplo, ventanas no solapadas) de cobertura. En algunas modalidades, la cobertura de todos los fragmentos puede incluir ventanas de fragmentos pequeños (por ejemplo, fragmentos de aproximadamente 100 pb a aproximadamente 150 pb de longitud). En algunas modalidades, la cobertura de todos los fragmentos puede incluir ventanas de fragmentos grandes (por ejemplo, fragmentos de aproximadamente 151 pb a aproximadamente 220 pb de longitud).

En ciertas modalidades, se puede usar un perfil de fragmentación del ADNlc para identificar los orígenes moleculares del ADNlc en los pacientes e identificar los atributos genómicos y cromatínicos asociados a los cambios de fragmentación.

En algunas modalidades, se puede usar un perfil de fragmentación del ADNlc para identificar el tejido de origen de un cáncer (por ejemplo, un cáncer hepático, un cáncer colorrectal, un cáncer de pulmón, un cáncer de mama, un cáncer gástrico, un cáncer de páncreas, un cáncer de vías biliares o un cáncer de ovario). Por ejemplo, se puede usar un perfil de fragmentación del ADNlc para identificar un cáncer localizado. Cuando un perfil de fragmentación del ADNlc incluye un perfil de región objetivo, una o más alteraciones descritas en la presente pueden usarse para identificar el tejido de origen de un cáncer. En algunas modalidades, una o más alteraciones en regiones cromosómicas pueden usarse para identificar el tejido de origen de un cáncer.

El perfil de fragmentación del ADNlc puede obtenerse usando cualquier método apropiado. En algunas modalidades, el ADNlc de un mamífero (por ejemplo, un mamífero que tiene o se sospecha que tiene cáncer) puede procesarse en bibliotecas de secuenciación que pueden someterse a secuenciación del genoma completo (por ejemplo, secuenciación del
5 genoma completo de baja cobertura), mapearse en el genoma y analizarse para determinar las longitudes de los fragmentos de ADNlc. Las secuencias mapeadas pueden analizarse en ventanas no solapadas que cubren el genoma. Las ventanas pueden ser de cualquier tamaño. Por ejemplo, las ventanas pueden tener de miles a millones de bases. Como ejemplo no limitante, una ventana puede tener una longitud aproximada de 5 megabases (Mb). Se puede
10 asignar cualquier número apropiado de ventanas. Por ejemplo, se pueden mapear de decenas a miles de ventanas en el genoma. Por ejemplo, se pueden mapear cientos o miles de ventanas en el genoma. Dentro de cada ventana puede determinarse un perfil de fragmentación del ADNlc.

En algunas modalidades, los métodos y materiales descritos en la presente también
15 pueden incluir el aprendizaje automático. Por ejemplo, se puede usar el aprendizaje automático para identificar un perfil de fragmentación alterado (por ejemplo, usando la cobertura de los fragmentos de ADNlc, el tamaño de los fragmentos de ADNlc, la cobertura de los cromosomas y el ADNmt).

En algunas modalidades, los métodos y materiales descritos en la presente pueden
20 ser el único método usado para identificar a un mamífero (por ejemplo, un humano) que tiene cáncer hepático. Por ejemplo, determinar un perfil de fragmentación del ADNlc puede ser el único método usado para identificar a un mamífero que tiene cáncer hepático.

En algunas modalidades, los métodos y materiales descritos en la presente pueden usarse junto con uno o más métodos adicionales para identificar a un mamífero (por ejemplo,
25 un humano) con cáncer hepático. Los ejemplos de métodos usados para identificar a un mamífero que tiene cáncer incluyen, sin limitación, identificar una o más alteraciones de secuencia específicas del cáncer, identificar una o más alteraciones cromosómicas (por ejemplo, aneuploidías y reordenamientos) e identificar otras alteraciones del ADNlc. Por ejemplo, determinar un perfil de fragmentación del ADNlc puede usarse junto con identificar
30 una o más mutaciones específicas del cáncer en el genoma de un mamífero para identificar a un mamífero que tiene cáncer hepático. Por ejemplo, determinar un perfil de fragmentación del ADNlc puede usarse junto con identificar una o más aneuploidías en el genoma de un mamífero para identificar a un mamífero que tiene cáncer hepático.

Métodos de tratamiento

Los métodos realizados en la presente incluyen identificar un mamífero que tiene cáncer. Los métodos incluyen, extraer ADN libre de células (ADNlc) de una muestra biológica del sujeto; generar bibliotecas genómicas a partir del ADNlc extraído; secuenciar moléculas individuales de ADNlc para obtener perfiles de fragmentación; comparar los perfiles de fragmentación con sujetos sanos y/o genoma de referencia, y diagnosticar si el sujeto tiene una enfermedad o trastorno hepático.

En ciertas modalidades, un método de diagnóstico del cáncer hepático comprende aislar el ADN libre de células (ADNlc) circulante de una muestra biológica y realizar la secuenciación del genoma completo de las moléculas de ADNlc para generar bibliotecas genómicas y perfiles de fragmentación; identificar los orígenes celulares de los perfiles de fragmentación del ADNlc que comprenden comparar los perfiles de fragmentación del genoma completo con la captura de conformación cromosómica de secuenciación de alto rendimiento (Hi-C); correlacionar los perfiles de fragmentación del ADNlc con la alteración de la unión al ADN de los factores de transcripción; determinar las ganancias o pérdidas cromosómicas en los sujetos con cáncer hepático en comparación con los sujetos sanos; ejecutar un modelo de aprendizaje automático para determinar los cambios en los perfiles de fragmentación del ADNlc que clasifica al sujeto como paciente con cáncer basado en el perfil de fragmentación del ADNlc para el sujeto; por lo mismo, diagnosticar el cáncer hepático y administrar un tratamiento contra el cáncer al sujeto.

En ciertas modalidades, a un sujeto se le diagnostica un cáncer, por ejemplo, en una etapa temprana. En ciertas modalidades, se identifica el tipo y la etapa del cáncer hepático y se trata al sujeto con una o más terapias contra el cáncer.

En algunas modalidades, los métodos y materiales descritos en la presente son para la valoración, monitorización y/o tratamiento de mamíferos (por ejemplo, humanos) que tienen o se sospecha que tienen cáncer. En algunas modalidades, se proporcionan métodos y materiales para identificar a un mamífero con cáncer hepático. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar si el mamífero tiene cáncer basándose, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas modalidades, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede valorarse para determinar el tejido de origen del cáncer en el mamífero basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero. En algunas modalidades, se proporcionan métodos y materiales para identificar a

un mamífero con cáncer hepático y administrar uno o más tratamientos al mamífero para tratar al mamífero. Por ejemplo, una muestra (por ejemplo, una muestra de sangre) obtenida de un mamífero puede ser valorada para determinar si el mamífero tiene cáncer hepático basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero, y administrar uno o más tratamientos contra el cáncer al mamífero. En algunas modalidades, se proporcionan métodos y materiales para tratar a un mamífero que tiene cáncer. Por ejemplo, uno o más tratamientos contra el cáncer pueden administrarse a un mamífero identificado que tiene cáncer (por ejemplo, basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero) para tratar al mamífero. En algunas modalidades, durante o después del curso de un tratamiento contra el cáncer (por ejemplo, cualquiera de los tratamientos contra el cáncer descritos en la presente), un mamífero puede someterse a monitorización (o ser seleccionado para una mayor monitorización) y/o a pruebas diagnósticas adicionales. En algunas modalidades, la monitorización puede incluir la valoración de mamíferos que tienen, o se sospecha que tienen, cáncer hepático, por ejemplo, al valorar una muestra (por ejemplo, una muestra de sangre) obtenida del mamífero para determinar el perfil de fragmentación del ADNlc del mamífero como se describe en la presente, y los cambios en los perfiles de fragmentación del ADNlc a lo largo del tiempo pueden usarse para identificar la respuesta al tratamiento y/o identificar al mamífero como portador de cáncer (por ejemplo, un cáncer residual).

Cualquier mamífero apropiado puede ser valorado, monitorizado y/o tratado como se describe en la presente. Un mamífero puede ser un mamífero con cáncer hepático. Puede tratarse de un mamífero sospechoso de tener cáncer hepático. Los ejemplos de mamíferos que pueden ser valorados, monitorizados y/o tratados como se describe en la presente incluyen, sin limitación, humanos, primates tales como monos, perros, gatos, caballos, vacas, cerdos, ovejas, ratones y ratas. Por ejemplo, un humano que tiene, o se sospecha que tiene, cáncer hepático puede ser valorado para determinar un perfil de fragmentación del ADNlc como se describe en la presente y, opcionalmente, puede ser tratado con uno o más tratamientos contra el cáncer como se describe en la presente.

Cualquier muestra apropiada de un mamífero puede ser valorada como se describe en la presente (por ejemplo, valorada para un patrón de fragmentación del ADN). En algunas modalidades, una muestra puede incluir ADN (por ejemplo, ADN genómico). En algunas modalidades, una muestra puede incluir ADNlc (por ejemplo, ADN tumoral circulante [ADNtc]). En algunas modalidades, una muestra puede ser una muestra líquida (por ejemplo, una biopsia líquida). Los ejemplos de muestras que pueden contener ADN y/o polipéptidos

incluyen, sin limitación, sangre (por ejemplo, sangre entera, suero o plasma), amnios, tejidos, orina, líquido cefalorraquídeo, saliva, esputo, lavado bronco-alveolar, bilis, líquido linfático, líquido de quiste, heces, ascitis, frotis de Papanicolaou, leche materna y condensado de aliento exhalado. Por ejemplo, puede valorarse una muestra de plasma para determinar un
5 perfil de fragmentación del ADNlc como se describe en la presente.

Una muestra de un mamífero que vaya a ser valorada como se describe en la presente (por ejemplo, valorada para un patrón de fragmentación del ADN) puede incluir cualquier cantidad apropiada del ADNlc. En algunas modalidades, una muestra puede incluir una cantidad limitante de ADN. Por ejemplo, puede obtenerse un perfil de fragmentación del ADNlc
10 a partir de una muestra que incluya menos ADN del que típicamente es necesario para otros métodos de análisis del ADNlc, tales como los descritos, por ejemplo, en Phallen *et al.*, 2017 *Sci Transl Med* 9; Cohen *et al.*, 2018 *Science* 359:926; Newman *et al.*, 2014 *Nat Med* 20:548; y Newman *et al.*, 2016 *Nat Biotechnol* 34:547).

En algunas modalidades, se puede procesar una muestra (por ejemplo, para aislar y/o
15 purificar ADN y/o polipéptidos de la muestra). Por ejemplo, aislar y/o purificar el ADN puede incluir la lisis celular (por ejemplo, usando detergentes y/o tensioactivos), eliminar proteínas (por ejemplo, usando una proteasa), y/o eliminar ARN (por ejemplo, usando una ARNasa). Como otro ejemplo, el aislamiento y/o purificación de polipéptidos puede incluir lisis celular (por ejemplo, usando detergentes y/o tensioactivos), eliminar ADN (por ejemplo, usando una
20 ADNasa), y/o eliminar ARN (por ejemplo, usando una ARNasa).

Un cáncer puede estar en cualquier etapa. En algunas modalidades, un cáncer puede ser un cáncer en etapa temprana. En algunas modalidades, un cáncer puede ser asintomático. En algunas modalidades, un cáncer puede ser una enfermedad residual y/o una recidiva (por ejemplo, después de una resección quirúrgica y/o después de una terapia contra el cáncer).
25 Un cáncer puede ser de cualquier tipo. Los ejemplos de tipos de cánceres que pueden ser valorados, monitorizados y/o tratados como se describe en la presente incluyen, sin limitación, cánceres colorrectales, cánceres de pulmón, cánceres de mama, cánceres gástricos, cánceres pancreáticos, cánceres de vías biliares y cánceres de ovario.

Cuando se trata a un mamífero que tiene, o se sospecha que tiene, cáncer hepático
30 como se describe en la presente, al mamífero se le puede administrar uno o más tratamientos contra el cáncer. Un tratamiento contra el cáncer puede ser cualquier tratamiento adecuado contra el cáncer. Uno o más tratamientos contra el cáncer descritos en la presente pueden administrarse a un mamífero con cualquier frecuencia adecuada (por ejemplo, una o varias

veces durante un periodo en el rango de días a semanas). Los ejemplos de tratamientos contra el cáncer incluyen, sin limitación, la quimioterapia adyuvante, la quimioterapia neoadyuvante, la radioterapia, la terapia hormonal, la terapia citotóxica, la inmunoterapia, la terapia con células T adoptivas (por ejemplo, receptores de antígenos quiméricos y/o células T con receptores de antígenos quiméricos o modificados), la terapia dirigida tal como la administración de inhibidores de la cinasa (por ejemplo, inhibidores de la cinasa dirigidos a una lesión genética particular, como una translocación o mutación), (por ejemplo, un inhibidor de la cinasa, un anticuerpo, un anticuerpo biespecífico), inhibidores de la transducción de señales, anticuerpos biespecíficos o fragmentos de anticuerpos (por ejemplo, BiTE, por sus siglas en inglés), anticuerpos monoclonales, inhibidores del punto de control inmunitario, cirugía (por ejemplo, resección quirúrgica), o cualquier combinación de los anteriores. En algunas modalidades, un tratamiento contra el cáncer puede reducir la gravedad del cáncer, reducir un síntoma del cáncer y/o reducir el número de células cancerosas presentes en el mamífero.

En algunas modalidades, un tratamiento contra el cáncer puede incluir un inhibidor del punto de control inmunitario. Los ejemplos no limitantes de inhibidores de puntos de control inmunitarios incluyen nivolumab (Opdivo), pembrolizumab (Keytruda), atezolizumab (tecentriq), avelumab (bavencio), durvalumab (imfinzi), ipilimumab (yervoy).

Las terapias contra el cáncer en general también incluyen una variedad de terapias combinadas con tratamientos químicos y de radiación. Las quimioterapias combinadas incluyen, por ejemplo, cisplatino (CDDP, por sus siglas en inglés), carboplatino, procarbazona, mecloretamina, ciclofosfamida, camptotecina, ifosfamida, melfalán, clorambucilo, busulfán, nitrosurea, dactinomicina, daunorrubicina, doxorubicina, bleomicina, plicomicina, mitomicina, etopósido (VP16), tamoxifeno, raloxifeno, agentes fijadores de receptores de estrógenos, taxol, gemcitabina, navelbina, inhibidores de la farnesil-proteína transferasa, transplatino, 5-fluorouracilo, vincristina, vinblastina y metotrexato, temazolomida (una forma acuosa de DTIC, por sus siglas en inglés), o cualquier variante análoga o derivada de las anteriores. La combinación de quimioterapia con terapia biológica se conoce como bioquimioterapia. La quimioterapia también puede administrarse en dosis bajas y continuas, lo que se conoce como quimioterapia metronómica.

Otras quimioterapias combinadas incluyen, por ejemplo, agentes alquilantes tales como tiotepa y ciclofosfamida; alquilsulfonatos tales como busulfán, improsulfán y piposulfán; aziridinas tales como benzodopa, carboquona, meturedopa y uredopa; etileniminas y

metilamelaminas, tales como altretamina, trietilenemelamina, trietilenfosforamida, trietilenotiofosforamida y trimetilolomelamina; acetogeninas (especialmente bullatacina y bullatacinona); una camptotecina (incluido el análogo sintético topotecán); briostatina; calilistatina; CC-1065 (incluidos sus análogos sintéticos adozelesina, carzelesina y bizelesina); 5
criptoficinas (particularmente criptoficina 1 y criptoficina 8); dolastatina; duocarmicina (incluidos los análogos sintéticos KW-2189 y CB1-TM1); eleuterobina; pancratistatina; una sarcodictina; espongistatina; mostazas nitrogenadas tales como clorambucilo, clorafanfazina, colofosfamida, estramustina, ifosfamida, mecloretamina, clorhidrato de óxido de mecloretamina, melfalán, novembichina, fenesterina, prednimustina, 10
trofosfamida, mostaza de uracilo; nitrosureas, tales como carmustina, clorozotocina, fotemustina, lomustina, nimustina y ranimustina; antibióticos, tales como los antibióticos de enediyina (por ejemplo, calicheamicina, especialmente calicheamicina gamma II y calicheamicina omega II; dinemicina, incluyendo dinemicina A; bisfosfonatos, tales como clodronato; una esperamicina; así como cromóforos neocarzinostatina y cromoproteína 15
relacionada a cromóforos antibióticos de enediyina, aclacinomisinas, actinomicina, autrarnycina, azaserina, bleomicinas, cactinomicina, carabicina, carminomicina, carzinofilina, cromomicina, dactinomicina, daunorrubicina, detorrubicina, 6-diazo-5-oxo-L-norleucina, doxorrubicina (incluida morfolino-doxorrubicina, cianomorfolino-doxorrubicina, 2-pirrolino-doxorrubicina y deoxidoxorrubicina), epirubicina, esorrubicina, idarrubicina, marcelomicina, 20
mitomicinas tales como la mitomicina C, ácido micofenólico, nogalarnicina, olivomicinas, peplomicina, potfiromicina, puromicina, quelamicina, rodorrubicina, estreptonigrina, estreptozocina, tubercidina, ubenimex, zinostatina, zorubicina; antimetabolitos como metotrexato y 5-fluorouracilo (5- FU); análogos del ácido fólico tales como denopterina, pteropterina, trimetrexato; análogos de las purinas tales como fludarabina, 6-mercaptapurina, 25
tiamiprina, tioguanina; análogos de las pirimidinas, tales como ancitabina, azacitidina, 6-azauridina, carmofur, citarabina, dideoxiuridina, doxifluridina, enocitabina, floxuridina; andrógenos, tales como calusterona, propionato de dromostanolona, epitioestano, mepitioestano, testolactona; antiadrenales, tales como mitotano, trilostano; reponedores del ácido fólico tales como ácido frolínico; aceglatona; glucósido de aldofosfamida; ácido 30
aminolevulínico; eniluracilo; amsacrina; bestrabucilo; bisantreno; edatraxato; defofamina; demecolcina; diaziquona; elformitina; acetato de eliptinio; una epotilona; etoglúcido; nitrato de galio; hidroxiiurea; lentinán; lonidainina; maytansinoides tales como maytansina y ansamitocinas; mitoguazona; mitoxantrona; mopidanmol; nitraerina; pentostatina; fenamet; pirarubicina; losoxantrona; ácido podofilínico; 2-etilhidrazida; procarbazona; complejo 35
polisacárido PSK (por sus siglas en inglés); razoxano; rizoxina; sizofirán; espirogermanio;

ácido tenuazónico; triaziquona; 2,2',2"-triclorotrietilamina; tricotecenos (especialmente toxina T-2, verracurina A, roridina A y anguidina); uretano; vindesina; dacarbazina; manomustina; mitobronitol; mitolactol; pipobromano; gacitosina; arabinósido ("Ara-C"); ciclofosfamida; taxoides, por ejemplo, paclitaxel y docetaxel gemcitabina; 6- tioguanina; mercaptopurina; 5 complejos de coordinación de platino, tales como cisplatino, oxaliplatino y carboplatino; vinblastina; platino etopósido (VP-16); ifosfamida; mitoxantrona; vincristina; vinorelbina; novantrona; tenipósido; edatrexato; daunomicina; aminopterina; xeloda; ibandronato; irinotecán (por ejemplo, CPT-11); inhibidor de la topoisomerasa RFS 2000; difluorometilornitina (DMFO, por sus siglas en inglés); retinoides como ácido retinoico; 10 capecitabina; carboplatino, procarbazona, plicomicina, gemcitabina, navelbina, inhibidores de la farnesil-proteína transferasa, transplatino; y sales, ácidos o derivados farmacéuticamente aceptables de cualquiera de los anteriores.

En general, los inmunoterapéuticos se basan en el uso de células y moléculas efectoras inmunitarias para atacar y destruir las células cancerosas. El efector inmunitario 15 puede ser, por ejemplo, un anticuerpo específico para algún marcador de la superficie de una célula tumoral. El anticuerpo por sí solo puede servir como efector de la terapia o puede reclutar a otras células para que realmente lleven a cabo la destrucción celular. El anticuerpo también puede conjugarse con un fármaco o toxina (quimioterapéutico, radionúclido, cadena A de la ricina, toxina del cólera, toxina de la tos ferina, etc.) y servir simplemente como agente 20 objetivo. Alternativamente, el efector puede ser un linfocito portador de una molécula de superficie que interactúa, directa o indirectamente, con un objetivo celular tumoral. Entre las células efectoras se encuentran las células T citotóxicas y las células NK (por sus siglas en inglés), así como variantes modificadas por ingeniería genética de estos tipos celulares para expresar receptores de antígeno quimérico.

25 La inmunoterapia puede comprender la supresión de células reguladoras T (Treg), células supresoras derivadas de mieloides (MDSC, por sus siglas en inglés) y fibroblastos asociados al cáncer (CAF, por sus siglas en inglés). En algunas modalidades, la inmunoterapia es una vacuna tumoral (por ejemplo, vacunas de células tumorales enteras, péptidos y vacunas recombinantes de antígenos asociados a tumores), o terapias celulares 30 adoptivas (ACT, por sus siglas en inglés) (por ejemplo, células T, células asesinas naturales, TIL y células LAK, por sus siglas en inglés, respectivamente). Las células T pueden estar diseñadas con receptores de antígeno quimérico (CAR, por sus siglas en inglés) o receptores de células T (TCR, por sus siglas en inglés) para antígenos tumorales específicos. Como se usa en la presente, un receptor de antígeno quimérico (o CAR) puede referirse a cualquier

receptor modificado por ingeniería específico para un antígeno de interés que, cuando se expresa en una célula T, confiere la especificidad del CAR a la célula T. Una vez creada usando técnicas moleculares estándar, una célula T que exprese un receptor de antígeno quimérico puede introducirse en un paciente, como se hace con una técnica como la transferencia celular adoptiva. En algunos aspectos, las células T son células T CD4 y/o CD8 activadas en el individuo que se caracterizan por ser células T CD4 y/o CD8 productoras de γ -IFN y/o potenciar la actividad citolítica en relación con antes de la administración de la combinación. Las células T CD4 y/o CD8 pueden presentar un aumento de la liberación de citocinas seleccionadas del grupo que consiste en IFN- γ , TNF- α e interleucinas. Las células T CD4 y/o CD8 pueden ser células T efectoras de memoria. En ciertas modalidades, las células T efectoras de memoria CD4 y/o CD8 se caracterizan por tener la expresión de CD44^{alta} CD62L^{baja}.

La inmunoterapia puede ser una vacuna contra el cáncer que comprende uno o más antígenos del cáncer, en particular una proteína o un fragmento inmunogénico de la misma, ADN o ARN que codifique dicho antígeno del cáncer, en particular una proteína o un fragmento inmunogénico de la misma, lisados de células cancerosas y/o preparaciones proteicas de células tumorales. Como se usa en la presente, un antígeno canceroso es una sustancia antigénica presente en células cancerosas. En principio, cualquier proteína producida en una célula cancerosa que tiene una estructura anormal debido a una mutación puede actuar como antígeno del cáncer. En principio, los antígenos del cáncer pueden ser productos de oncogenes y genes supresores de tumores mutados, productos de otros genes mutados, proteínas celulares sobreexpresadas o expresadas de forma aberrante, antígenos del cáncer producidos por virus oncogénicos, antígenos oncofetales, glucolípidos y glucoproteínas de la superficie celular alterados o antígenos de diferenciación específicos de un tipo celular. Los ejemplos de antígenos del cáncer incluyen los productos anómalos de los genes ras y p53. Otros ejemplos incluyen antígenos de diferenciación tisular, antígenos de proteínas mutantes, antígenos virales oncogénicos, antígenos de cáncer-testis y antígenos específicos vasculares o estromales. Los antígenos de diferenciación tisular son aquellos que son específicos de un cierto tipo de tejido.

La inmunoterapia puede ser un anticuerpo, tal como parte de una preparación de anticuerpos policlonales, o puede ser un anticuerpo monoclonal. El anticuerpo puede ser un anticuerpo humanizado, un anticuerpo quimérico, un fragmento de anticuerpo, un anticuerpo biespecífico o un anticuerpo de cadena simple. Un anticuerpo como se divulga en la presente incluye un fragmento de anticuerpo, tal como, pero no se limita a, Fab, Fab' y F(ab')₂, Fd, Fv

de cadena única (scFv), anticuerpos de cadena única, Fv con enlace disulfuro (sdfv) y fragmentos que incluyen un dominio VL o VH (por sus siglas en inglés, respectivamente). En algunos aspectos, el anticuerpo o fragmento del mismo se une específicamente al receptor del factor de crecimiento epidérmico (EGFR1, Erb-B1, por sus siglas en inglés), HER2/neu (Erb-B2), CD20, factor de crecimiento endotelial vascular (VEGF, por sus siglas en inglés), receptor del factor de crecimiento similar a la insulina (IGF-1R, por sus siglas en inglés), receptor TRAIL (por sus siglas en inglés), molécula de adhesión de células epiteliales, antígeno carcinoembrionario, antígeno de membrana específico de la próstata, mucina-1, CD30, CD33 o CD40.

10 Los ejemplos de anticuerpos monoclonales incluyen, sin limitación, trastuzumab (anticuerpo anti-HER2/neu); Pertuzumab (mAb anti-HER2); cetuximab (anticuerpo monoclonal quimérico contra el receptor del factor de crecimiento epidérmico EGFR); panitumumab (anticuerpo anti-EGFR); nimotuzumab (anticuerpo anti-EGFR); Zalutumumab (mAb anti-EGFR); Necitumumab (mAb anti-EGFR); MDX-210 (anticuerpo biespecífico humanizado anti-HER-2); MDX-210 (anticuerpo biespecífico humanizado anti-HER-2); MDX-447 (anticuerpo biespecífico humanizado anti-receptor de EGF); Rituximab (mAb quimérico murino/humano anti-CD20); Obinutuzumab (mAb anti-CD20); Ofatumumab (mAb anti-CD20); Tositumumab-I131 (mAb anti-CD20); Ibritumomab tiuxetán (mAb anti-CD20); Bevacizumab (mAb anti-VEGF); Ramucirumab (mAb anti-VEGFR2); Ranibizumab (mAb anti-VEGF); Aflibercept (dominios extracelulares de VEGFR1 y VEGFR2 fusionados a IgG1 Fc); AMG386 (péptido de unión de angiopoyetina-1 y -2 fusionado a IgG1 Fc); Dalotuzumab (mAb anti-IGF-1R); Gemtuzumab ozogamicina (mAb anti-CD33); Alemtuzumab (mAb anti-Campath-1/CD52); Brentuximab vedotina (mAb anti-CD30); Catumaxomab (mAb biespecífico dirigido a la molécula de adhesión de células epiteliales y CD3); Naptumomab (mAb anti-5T4); Girentuximab (anti-Anhidrasa carbónica ix); o Farletuzumab (anti-receptor de folato). Otros ejemplos incluyen anticuerpos tales como Panorex.TM. (17-1A) (anticuerpo monoclonal murino); Panorex (MAB17-1A) (anticuerpo monoclonal murino quimérico); BEC2 (amidiotípico mAb, imita el epítipo GD) (con BCG); Oncolym (anticuerpo monoclonal Lym-1); SMART M195 Ab, humanizado 13' 1 LYM-1 (Oncolym), Ovarex (B43.13, mAb anti-idiotípico de ratón); 3622W94 mAb que se une al antígeno de pancarcinoma EGP40 (17-1A) en adenocarcinomas; Zenapax (SMART Anti-Tac (receptor de IL-2); SMART M195 Ab, humanizado Ab, humanizado); NovoMAb-G2 (Ab específico de pancarcinoma); TNT (mAb quimérico frente a antígenos histónicos); TNT (mAb quimérico frente a antígenos histónicos); Gliomab-H (Ab monoclonales-humanizados); GNI-250 Mab; EMD-72000 (antagonista quimérico-EGF); LymphoCide (anticuerpo IL.L.2); y MDX-260 biespecífico, objetivos GD-2,

ANA Ab, SMART IDIO Ab, SMART ABL 364 Ab o ImmuRAIT-CEA (por sus siglas y nombres en inglés, respectivamente). Otros ejemplos de anticuerpos incluyen Zanolimumab (mAb anti-CD4), Keliximab (mAb anti-CD4); Ipilimumab (MDX-101; mAb anti-CTLA-4); Tremilimumab (mAb anti-CTLA-4); (Daclizumab (mAb anti-CD25/IL-2R); Basiliximab (mAb anti-CD25/IL-2R);

5 MDX-1106 (mAb anti PD1); anticuerpo contra el GITR; GC1008 (anticuerpo anti-TGF- β); metelimumab/CAT-192 (anticuerpo anti-TGF- β); Ierdelimumab/CAT-152 (anticuerpo anti-TGF- β); ID11 (anticuerpo anti-TGF- β); Denosumab (mAb anti-RANKL); BMS-663513 (mAb humanizado anti-4-1BB); SGN-40 (mAb humanizado anti-CD40); CP870,893 (mAb anti-CD40 humano); Infliximab (mAb anti-TNF quimérico; Adalimumab (mAb anti-TNF humano);

10 Certolizumab (Fab anti-TNF humanizado); Golimumab (anti-TNF); Etanercept (dominio extracelular de TNFR fusionado a IgG1 Fc); Belatacept (dominio extracelular de CTLA-4 fusionado a Fc); Abatacept (dominio extracelular de CTLA-4 fusionado a Fc); Belimumab (estimulador de linfocitos anti-B); Muromonab-CD3 (mAb anti-CD3); Otelixizumab (mAb anti-CD3); Teplizumab (mAb anti-CD3); Tocilizumab (mAb anti-IL6R); REGN88 (mAb anti-IL6R);

15 Ustekinumab (anti-IL-12/23 mAb); Briakinumab (anti-IL-12/23 mAb); Natalizumab (anti- α 4 integrina); Vedolizumab (anti- α 4 β 7 integrina mAb); T1 h (anti- CD6 mAb); Epratuzumab (mAb anti-CD22); Efalizumab (mAb anti-CD11a); y Atacicept (dominio extracelular del activador transmembranal e interaccionador del ligando modulador del calcio fusionado con Fc); (por sus siglas y nombres en inglés, respectivamente).

20 Cuando se monitoriza a un mamífero que tiene, o se sospecha que tiene, cáncer como se describe en la presente (por ejemplo, basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero), la monitorización puede realizarse antes, durante y/o después del curso de un tratamiento contra el cáncer. Los métodos de monitorización proporcionados en la presente pueden usarse para determinar la eficacia de uno o más

25 tratamientos contra el cáncer y/o para seleccionar un mamífero para una mayor monitorización. En algunas modalidades, la monitorización puede incluir identificar un perfil de fragmentación del ADNlc como se describe en la presente. Por ejemplo, puede obtenerse un perfil de fragmentación del ADNlc antes de administrar uno o más tratamientos contra el cáncer a un mamífero que tiene, o se sospecha que tiene, cáncer, pueden administrarse uno

30 o más tratamientos contra el cáncer al mamífero, y pueden obtenerse uno o más perfiles de fragmentación del ADNlc durante el curso del tratamiento contra el cáncer. En algunas modalidades, un perfil de fragmentación del ADNlc puede cambiar durante el curso del tratamiento del cáncer (por ejemplo, cualquiera de los tratamientos del cáncer descritos en la presente). Por ejemplo, un perfil de fragmentación del ADNlc indicativo de que el mamífero

35 tiene cáncer puede cambiar a un perfil de fragmentación del ADNlc indicativo de que el

mamífero no tiene cáncer. Tal cambio en el perfil de fragmentación del ADNlc puede indicar que el tratamiento del cáncer está funcionando. Por el contrario, un perfil de fragmentación del ADNlc puede permanecer estático (por ejemplo, igual o aproximadamente igual) durante el curso del tratamiento del cáncer (por ejemplo, cualquiera de los tratamientos del cáncer descritos en la presente). Tal perfil estático de fragmentación del ADNlc puede indicar que el tratamiento del cáncer no está funcionando.

En algunas modalidades, la monitorización puede incluir técnicas convencionales capaces de monitorizar uno o más tratamientos contra el cáncer (por ejemplo, la eficacia de uno o más tratamientos contra el cáncer). En algunas modalidades, a un mamífero seleccionado para una mayor vigilancia se le puede administrar una prueba diagnóstica (por ejemplo, cualquiera de las pruebas diagnósticas divulgadas en la presente) con una mayor frecuencia en comparación con un mamífero que no ha sido seleccionado para una mayor vigilancia. Por ejemplo, a un mamífero seleccionado para una mayor monitorización se le puede administrar una prueba diagnóstica con una frecuencia de dos veces al día, diaria, quincenal, semanal, bimensual, mensual, trimestral, semestral, anual, o cualquier frecuencia dentro de las mismas. En algunas modalidades, a un mamífero seleccionado para una mayor monitorización se le puede administrar una o más pruebas diagnósticas adicionales en comparación con un mamífero que no ha sido seleccionado para una mayor monitorización. Por ejemplo, a un mamífero seleccionado para una mayor monitorización se le pueden administrar dos pruebas diagnósticas, mientras que a un mamífero que no ha sido seleccionado para una mayor monitorización solamente se le administra una única prueba diagnóstica (o ninguna prueba diagnóstica). En algunas modalidades, un mamífero que ha sido seleccionado para una mayor monitorización también puede ser seleccionado para otras pruebas de diagnóstico. Una vez que se ha identificado la presencia de un tumor o un cáncer (por ejemplo, una célula cancerosa) (por ejemplo, mediante cualquiera de los diversos métodos divulgados en la presente), puede ser beneficioso que el mamífero se someta tanto a una mayor monitorización (por ejemplo, para valorar la progresión del tumor o cáncer en el mamífero y/o valorar el desarrollo de uno o más biomarcadores de cáncer tales como mutaciones), y más pruebas de diagnóstico (por ejemplo, para determinar el tamaño y/o localización exacta (por ejemplo, tejido de origen) del tumor o cáncer). En algunas modalidades, pueden administrarse uno o más tratamientos contra el cáncer al mamífero seleccionado para una mayor monitorización después de que se detecte un biomarcador de cáncer y/o después de que el perfil de fragmentación del ADNlc del mamífero no haya mejorado o se haya deteriorado. Puede administrarse cualquiera de los tratamientos contra el cáncer divulgados en la presente o conocidos en la materia. Por ejemplo, un mamífero que

ha sido seleccionado para una mayor monitorización puede ser sometido a una monitorización adicional, y puede administrarse un tratamiento contra el cáncer si la presencia de la célula cancerosa se mantiene a lo largo del periodo de mayor monitorización. Adicional o alternativamente, a un mamífero que ha sido seleccionado para una mayor monitorización se le puede administrar un tratamiento contra el cáncer, y monitorizarlo aún más a medida que progresa el tratamiento contra el cáncer. En algunas modalidades, después de administrar un tratamiento contra el cáncer a un mamífero que ha sido seleccionado para una mayor monitorización, la mayor monitorización revelará uno o más biomarcadores de cáncer (por ejemplo, mutaciones). En algunas modalidades, tal uno o más biomarcadores del cáncer proporcionarán la causa para administrar un tratamiento diferente del cáncer (por ejemplo, una mutación de resistencia puede surgir en una célula cancerosa durante el tratamiento del cáncer, que la célula cancerosa que alberga la mutación de resistencia es resistente al tratamiento original del cáncer).

Cuando se identifica que un mamífero tiene cáncer como se describe en la presente (por ejemplo, basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero), la identificación puede ser antes y/o durante el curso de un tratamiento contra el cáncer. Los métodos para identificar a un mamífero que tiene cáncer que se proporcionan en la presente pueden usarse como primer diagnóstico para identificar al mamífero (por ejemplo, que tiene cáncer antes de cualquier tratamiento) y/o para seleccionar al mamífero para otras pruebas diagnósticas. En algunas modalidades, una vez que se ha determinado que un mamífero tiene cáncer, se le pueden administrar otras pruebas y/o seleccionarlo para otras pruebas diagnósticas. En algunas modalidades, los métodos proporcionados en la presente pueden usarse para seleccionar un mamífero para pruebas diagnósticas adicionales en un periodo anterior al periodo en el que las técnicas convencionales son capaces de diagnosticar al mamífero con un cáncer en etapa temprana. Por ejemplo, los métodos proporcionados en la presente para seleccionar un mamífero para más pruebas diagnósticas pueden usarse cuando un mamífero no ha sido diagnosticado de cáncer por métodos convencionales y/o cuando no se sabe que un mamífero alberga un cáncer. En algunas modalidades, a un mamífero seleccionado para pruebas diagnósticas posteriores se le puede administrar una prueba diagnóstica (por ejemplo, cualquiera de las pruebas diagnósticas divulgadas en la presente) con una frecuencia mayor en comparación con un mamífero que no ha sido seleccionado para pruebas diagnósticas posteriores. Por ejemplo, a un mamífero seleccionado para una prueba diagnóstica posterior se le puede administrar una prueba diagnóstica con una frecuencia de dos veces al día, diaria, quincenal, semanal, bimensual, mensual, trimestral, semestral, anual, o con cualquier frecuencia dentro de las mismas. En

algunas modalidades, a un mamífero seleccionado para pruebas diagnósticas adicionales se le puede administrar una o más pruebas diagnósticas adicionales en comparación con un mamífero que no ha sido seleccionado para pruebas diagnósticas adicionales. Por ejemplo, a un mamífero seleccionado para pruebas diagnósticas adicionales se le pueden administrar dos pruebas diagnósticas, mientras que a un mamífero que no ha sido seleccionado para pruebas diagnósticas adicionales se le administra una sola prueba diagnóstica (o ninguna prueba diagnóstica). En algunas modalidades, el método de prueba diagnóstica puede determinar la presencia del mismo tipo de cáncer (por ejemplo, que tiene el mismo tejido u origen) que el cáncer que se detectó originalmente (por ejemplo, basado, al menos en parte, en el perfil de fragmentación del ADNlc del mamífero). Adicional o alternatively, el método de prueba diagnóstica puede determinar la presencia de un tipo de cáncer diferente al detectado originalmente. En algunas modalidades, el método de prueba diagnóstica es una exploración. En algunas modalidades, la exploración es una tomografía computarizada (TC), una angiografía por TC (ATC), un esofagrama (una deglución de bario), un enema de bario, una resonancia magnética (RM), una PET, una ecografía (por ejemplo, una ecografía endobronquial, una ecografía endoscópica), una radiografía, una DEXA; (por sus siglas en inglés, respectivamente).

En algunas modalidades, el método de prueba diagnóstica es un examen físico, tal como una anoscopia, una broncoscopia (por ejemplo, una broncoscopia de autofluorescencia, una broncoscopia de luz blanca, una broncoscopia de navegación), una colonoscopia, una tomosíntesis digital de mama, una colangiopancreatografía retrógrada endoscópica (CPRE, por sus siglas en inglés), una esofagogastroduodenoscopia, una mamografía, una citología vaginal, un examen pélvico, una tomografía por emisión de positrones y tomografía computarizada (PET-CT). En algunas modalidades, un mamífero que ha sido seleccionado para otras pruebas diagnósticas también puede ser seleccionado para una mayor monitorización. Una vez que se ha identificado la presencia de un tumor o un cáncer (por ejemplo, una célula cancerosa) (por ejemplo, por cualquiera de los diversos métodos divulgados en la presente), puede ser beneficioso que el mamífero se someta tanto a una mayor monitorización (por ejemplo, para valorar la progresión del tumor o del cáncer en el mamífero y/o para valorar el desarrollo de uno o más biomarcadores del cáncer tales como mutaciones), como a más pruebas diagnósticas (por ejemplo, para determinar el tamaño y/o la localización exacta del tumor o del cáncer). En algunas modalidades, se administra un tratamiento contra el cáncer al mamífero seleccionado para realizar más pruebas diagnósticas después de que se detecte un biomarcador de cáncer y/o después de que el perfil de fragmentación del ADNlc del mamífero no haya mejorado o se haya deteriorado. Puede

administrarse cualquiera de los tratamientos contra el cáncer divulgados en la presente o conocidos en la materia. Por ejemplo, a un mamífero que ha sido seleccionado para una prueba diagnóstica adicional se le puede administrar una prueba diagnóstica adicional, y se le puede administrar un tratamiento contra el cáncer si se confirma la presencia del tumor o del cáncer. Adicional o alternativamente, a un mamífero que ha sido seleccionado para pruebas diagnósticas adicionales se le puede administrar un tratamiento contra el cáncer, y se le puede seguir monitorizando a medida que progresa el tratamiento contra el cáncer. En algunas modalidades, después de administrar un tratamiento contra el cáncer a un mamífero que ha sido seleccionado para pruebas diagnósticas adicionales, las pruebas adicionales revelarán uno o más biomarcadores del cáncer. En algunas modalidades, tal uno o más biomarcadores del cáncer (por ejemplo, mutaciones) proporcionarán la causa para administrar un tratamiento diferente del cáncer (por ejemplo, una mutación de resistencia puede surgir en una célula cancerosa durante el tratamiento del cáncer, que la célula cancerosa que alberga la mutación de resistencia es resistente al tratamiento original del cáncer).

15 **Sistemas**

En algunos ejemplos, la presente divulgación proporciona sistemas, métodos o kits que pueden incluir análisis de datos realizados en dispositivos de medición (por ejemplo, instrumentos de laboratorio, tales como una máquina de secuenciación), código de software que se ejecuta en hardware informático. El software puede almacenarse en la memoria y ejecutarse en uno o varios procesadores de hardware. El software puede organizarse en rutinas o paquetes que pueden comunicarse entre sí. Un módulo puede comprender uno o más dispositivos/computadoras, y potencialmente una o más rutinas/paquetes de software que se ejecutan en el uno o más dispositivos/computadoras. Por ejemplo, una aplicación o sistema de análisis puede incluir al menos un módulo de recepción de datos, un módulo de preprocesamiento de datos, un módulo de análisis de datos (que puede operar sobre uno o más tipos de datos genómicos), un módulo de interpretación de datos o un módulo de visualización de datos.

El módulo de recepción de datos puede conectar el hardware o la instrumentación del laboratorio con los sistemas informáticos que procesan los datos del laboratorio. El módulo de preprocesamiento de datos puede operar con los datos para prepararlos para el análisis. Los ejemplos de operaciones que pueden aplicarse a los datos en el módulo de preprocesamiento incluyen las transformaciones afines, las operaciones de eliminación de ruido, la limpieza de datos, el reformato o el submuestreo. El módulo de análisis de datos, que puede estar

especializado para analizar datos genómicos de uno o más materiales genómicos, puede, por ejemplo, tomar secuencias genómicas ensambladas y realizar análisis probabilísticos y estadísticos para identificar patrones anormales relacionados con una enfermedad, patología, estado, riesgo, condición o fenotipo. El módulo de interpretación de datos puede usar métodos de análisis, por ejemplo, extraídos de la estadística, las matemáticas o la biología, para apoyar la comprensión de la relación entre los patrones anormales identificados y las condiciones de salud, los estados funcionales, los pronósticos o los riesgos. El módulo de análisis de datos y/o el módulo de interpretación de datos pueden incluir uno o más modelos de aprendizaje automático, que pueden implementarse en hardware, por ejemplo, que ejecute software que incorpore un modelo de aprendizaje automático. El módulo de visualización de datos puede usar métodos de modelización matemática, infografía o renderización para crear representaciones visuales de los datos que puedan facilitar la comprensión o interpretación de los resultados. La presente divulgación proporciona sistemas de computadora que se programan para implementar métodos de la divulgación.

En algunas modalidades, los métodos divulgados en la presente pueden incluir análisis computacionales sobre datos de secuenciación de ácidos nucleicos de muestras de un individuo o de una pluralidad de individuos. Un análisis puede identificar una variante inferida a partir de datos de secuencia para identificar variantes de secuencia basadas en modelado probabilístico, modelado estadístico, modelado mecanístico, modelado de red o inferencias estadísticas. Los ejemplos no limitantes de métodos de análisis incluyen el análisis de componentes principales, los autocodificadores, la descomposición de valores singulares, las bases de Fourier, las ondículas, el análisis discriminante, la regresión, las máquinas de vectores soporte, los métodos basados en árboles, las redes, la factorización de matrices y la agrupación. Los ejemplos no limitantes de variantes incluyen una variación de la línea germinal o una mutación somática. En algunos ejemplos, una variante puede referirse a una variante ya conocida. La variante ya conocida puede confirmarse científicamente o reportarse en la literatura. En algunos ejemplos, una variante puede referirse a una variante putativa asociada a un cambio biológico. Un cambio biológico puede ser conocido o desconocido. En algunos ejemplos, una variante putativa puede estar descrita en la literatura, pero aún no confirmada biológicamente. Alternativamente, una variante putativa nunca se ha publicado en la literatura, pero puede inferirse basada en un análisis computacional divulgado en la presente. En algunos ejemplos, las variantes de la línea germinal pueden referirse a ácidos nucleicos que inducen variaciones naturales o normales.

En ciertas modalidades, el sistema informático incluye una unidad central de

procesamiento (CPU, por sus siglas en inglés, también “procesador” y “procesador informático” en la presente), que puede ser un procesador de núcleo único o de núcleo múltiple, o una pluralidad de procesadores para el procesamiento en paralelo; memoria (por ejemplo, caché, memoria de acceso aleatorio, memoria de solamente lectura, memoria flash u otra memoria); unidad de almacenamiento electrónico (por ejemplo, disco duro), interfaz de comunicación (por ejemplo, adaptador de red) para comunicarse con uno o más sistemas; y dispositivos periféricos, tales como adaptadores para caché, otra memoria, almacenamiento de datos y/o pantalla electrónica. La memoria, la unidad de almacenamiento, la interfase y los dispositivos periféricos pueden estar en comunicación con la CPU a través de un bus de comunicación (líneas continuas), tal como una placa base. La unidad de almacenamiento puede ser una unidad de almacenamiento de datos (o repositorio de datos) para almacenar datos. Se pueden introducir uno o más atributos del analito desde uno o más dispositivos de medición. En la presente se describen ejemplos de analitos y dispositivos de medición.

El sistema informático puede acoplarse operativamente a una red informática (“red”) con ayuda de la interfase de comunicación. La red puede ser Internet, una intranet y/o extranet, o una intranet y/o extranet que esté en comunicación con Internet. En algunos casos, la red es una red de telecomunicaciones y/o de datos. La red puede incluir uno o más servidores informáticos, que pueden permitir la computación distribuida, tal como la computación en nube a través de la red (“la nube”) para realizar diversos aspectos del análisis, cálculo y generación de la presente divulgación, tales como, por ejemplo, la activación de una válvula o bomba para transferir un reactivo o muestra de una cámara a otra o la aplicación de calor a una muestra (por ejemplo, durante una reacción de amplificación), otros aspectos del procesamiento y/o ensayo de una muestra, la modalidad de análisis de secuenciación, la medición de conjuntos de valores representativos de clases de moléculas, identificar conjuntos de atributos y vectores de atributos a partir de datos de ensayo, el procesamiento de vectores de atributos usando un modelo de aprendizaje automático para obtener clasificaciones de salida, y el entrenamiento de un modelo de aprendizaje automático (por ejemplo, la búsqueda iterativa de valores óptimos de los parámetros del modelo de aprendizaje automático). Dicha computación en nube puede ser proporcionada por plataformas de computación en nube tales como, por ejemplo, Servicios Web de Amazon (AWS, por sus siglas en inglés), Microsoft Azure, plataforma en la nube de Google y nube de IBM. La red, en algunos casos con la ayuda del sistema informático, puede implementar una red par-a-par, que puede permitir a los dispositivos acoplados al sistema informático comportarse como un cliente o un servidor.

La CPU puede ejecutar una secuencia de instrucciones legibles por máquina, que pueden estar incorporadas en un programa o software. Las instrucciones pueden almacenarse en una ubicación de memoria, tal como la memoria. Las instrucciones pueden dirigirse a la CPU, que subsecuentemente puede programar o configurar de otro modo la CPU para implementar métodos de la presente divulgación. La CPU puede formar parte de un circuito, tal como un circuito integrado. Uno o más componentes del sistema pueden incluirse en el circuito. En algunos casos, el circuito es un circuito integrado de aplicación específica (ASIC, por sus siglas en inglés).

La unidad de almacenamiento puede guardar archivos, tales como controladores, bibliotecas y programas guardados. La unidad de almacenamiento puede almacenar datos del usuario, por ejemplo, preferencias y programas del usuario. En algunos casos, el sistema informático puede incluir una o más unidades adicionales de almacenamiento de datos externas al sistema informático, tales como las ubicadas en un servidor remoto que esté en comunicación con el sistema informático a través de una intranet o de Internet.

El sistema informático puede comunicarse con uno o más sistemas informáticos remotos a través de la red. Por ejemplo, el sistema informático puede comunicarse con un sistema informático remoto de un usuario. Los ejemplos de sistemas informáticos remotos incluyen computadoras personales (por ejemplo, portátiles), tabletas (por ejemplo, Apple® iPad, Samsung® Galaxy Tab), teléfonos, teléfonos inteligentes (por ejemplo, Apple® iPhone, dispositivos con Android, Blackberry®) o asistentes personales digitales. El usuario puede acceder al sistema informático mediante la red.

Los métodos descritos en la presente pueden implementarse mediante código ejecutable por máquina (por ejemplo, procesador de computadora) almacenado en una ubicación de almacenamiento electrónico del sistema informático tal como, por ejemplo, en la memoria o unidad de almacenamiento electrónico. El código ejecutable o legible por máquina puede proporcionarse en forma de software. Durante su uso, el código puede ser ejecutado por la CPU. En algunos casos, el código puede recuperarse de la unidad de almacenamiento y almacenarse en la memoria para que la CPU pueda acceder al mismo. En algunas situaciones, puede excluirse la unidad de almacenamiento electrónico, y las instrucciones ejecutables por máquina se almacenan en la memoria.

El código puede precompilarse y configurarse para usarlo con una máquina que tiene un procesador adaptado para ejecutar el código o puede compilarse durante el tiempo de ejecución. El código puede suministrarse en un lenguaje de programación que puede

seleccionarse para permitir que el código se ejecute de forma precompilada o como compilado.

Aspectos de los sistemas y métodos proporcionados en la presente, tales como el sistema informático, pueden llevarse a cabo mediante programación. Varios aspectos de la tecnología pueden considerarse “productos” o “artículos de fabricación” típicamente en forma de código ejecutable por máquina (o procesador) y/o datos asociados que se transportan o incorporan en un tipo de medio legible por máquina. El código ejecutable por máquina se puede almacenar en una unidad de almacenamiento electrónico, tal como una memoria (por ejemplo, memoria de solo lectura, memoria de acceso aleatorio, memoria flash) o un disco duro. Los medios de tipo “almacenamiento” pueden incluir cualquiera o toda la memoria tangible de las computadoras, procesadores o similares, o módulos asociados de estos, tal como varias memorias de semiconductores, unidades de cinta, unidades de disco y similares, que pueden proporcionar almacenamiento no transitorio en cualquier momento para la programación de software. Todo o porciones del software se pueden comunicar a veces a través de Internet u otras redes de telecomunicaciones. Estas comunicaciones, por ejemplo, pueden permitir la carga del software de una computadora o procesador en otro, por ejemplo, de un servidor de gestión o computadora anfitrión en la plataforma de computadora de un servidor de aplicaciones. Así, otro tipo de medios que pueden portar los elementos de software incluyen las ondas ópticas, eléctricas y electromagnéticas, como se usan a través de interfaces físicas entre dispositivos locales, a través de redes fijas cableadas y ópticas y a través de diversos enlaces aéreos. Los elementos físicos que transportan tales ondas, tales como enlaces por cable o inalámbricos, enlaces ópticos o similares, también pueden considerarse medios portadores del software. Como se usa en la presente, a menos que se restrinja a medios de “almacenamiento” tangibles y no transitorios, los términos tales como “medio legible por computadora o máquina” se refieren a cualquier medio que participa para proporcionar instrucciones a un procesador para su ejecución.

Por lo tanto, un medio legible por máquina, tal como un código ejecutable por computadora, puede adoptar muchas formas, entre las que se incluyen, pero no se limitan a, 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 de cualquier computadora(es) o como se puede usar para implementar las bases de datos, etc. mostradas en los dibujos. Los medios de almacenamiento volátiles incluyen memoria dinámica, tal como una memoria principal de esta plataforma de computadora. Los medios de transmisión

tangibles incluyen cables coaxiales; alambres de cobre y fibra óptica, que incluyen los alambres que comprenden un bus dentro de un sistema de computadora.

Los medios de transmisión de ondas portadoras pueden adoptar la forma de señales eléctricas o electromagnéticas, u ondas acústicas o de luz tal como aquellas generadas durante comunicaciones de datos por radiofrecuencia (RF, por sus siglas en inglés) e infrarrojo (IR, por sus siglas en inglés). Por lo tanto, las formas comunes de medios legibles por computadora 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, cinta de papel de tarjetas perforadas, cualquier otro medio de almacenamiento físico con patrones de agujeros, una RAM, una ROM, una PROM y EPROM, una FLASH-EPROM, cualquier otro chip o cartucho de memoria, una onda portadora que transporta datos o instrucciones, cables o enlaces que transportan esta onda portadora, o cualquier otro medio desde el cual una computadora puede leer código de programación y/o datos; (por sus siglas en inglés, respectivamente). Muchas de estas formas de medios legibles por computadora pueden participar en el transporte de una o más secuencias de una o más instrucciones a un procesador para su ejecución.

El sistema informático puede incluir o estar en comunicación con una pantalla electrónica que comprende una interfaz de usuario (UI, por sus siglas en inglés) para proporcionar, por ejemplo, una etapa actual de procesamiento o ensayo de una muestra (por ejemplo, una etapa particular, tal como un paso de lisis, o un paso de secuenciación que se está realizando). El sistema informático recibe las entradas de una o varias mediciones. Los ejemplos de interfaces de usuario incluyen, sin limitación, una interfaz gráfica de usuario (GUI, por sus siglas en inglés) y una interfaz de usuario basada en web. El algoritmo puede, por ejemplo, procesar y/o ensayar una muestra, realizar análisis de secuenciación, medir conjuntos de valores representativos de clases de moléculas, identificar conjuntos de atributos y vectores de atributos a partir de datos de ensayo, procesar vectores de atributos usando un modelo de aprendizaje automático para obtener clasificaciones de salida, y entrenar un modelo de aprendizaje automático (por ejemplo, buscar iterativamente los valores óptimos de los parámetros del modelo de aprendizaje automático).

En algunas modalidades, los sistemas capaces de ejecutar uno o más algoritmos, por ejemplo, computadoras portátiles, computadoras de sobremesa, iPads, dispositivos móviles, etc., para determinar los cambios en los perfiles de fragmentación del ADNlc clasifican al sujeto como paciente con cáncer basándose en el perfil de fragmentación del ADNlc para el

sujeto. Estos sistemas ejecutan además algoritmos de aprendizaje automático que pueden usarse para generar modelos tales como, por ejemplo, poblaciones de alto riesgo y poblaciones generales de bajo riesgo (una regresión logística penalizada con el método Mathios *et al.* (Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 2021;12(1):5060) atributos, así como la cobertura de los sitios de unión a factores de transcripción. Estos modelos pueden ser entrenados en la cohorte de sujetos con validación cruzada de 5 veces con 10 repeticiones, y las puntuaciones para cada muestra se calculan por la media a través de repeticiones y se evalúan usando ABC-ROC. Por ejemplo, el primer modelo usó los pacientes de alto riesgo sin cáncer y con HCC, mientras que el segundo usó los individuos sin cáncer y sin patología hepática. El modelo de alto riesgo bloqueado entrenado en la cohorte se aplicó a una segunda cohorte diferente para generar predicciones de cáncer en un conjunto de validación externo. A cada muestra se le puede aplicar una “etiqueta de clase” que indica la clasificación de la muestra para cualquier número de atributos de entrada. Por ejemplo, las etiquetas de clase para el conjunto de cohortes podrían indicar la identidad de los perfiles de fragmentación del ADNlc basados en la localización genómica, etc. Los conjuntos de entrenamiento resultantes se proporcionan a la unidad de aprendizaje automático, tal como una red neuronal o una máquina de vectores de soporte. Usando el conjunto de entrenamiento, la unidad de aprendizaje automático puede generar un modelo para clasificar la muestra de acuerdo con el perfil de fragmentación del ADNlc.

En algunas modalidades, se proporciona un método para crear un clasificador entrenado, que comprende los pasos de: (a) proporcionar una pluralidad de clases diferentes, en donde cada clase representa un conjunto de sujetos con una característica compartida (por ejemplo, de una o más cohortes); (b) proporcionar un modelo multiparamétrico representativo de las moléculas de ADN libres de células de cada una de una pluralidad de muestras pertenecientes a cada una de las clases, proporcionando así un conjunto de datos de entrenamiento; y (c) entrenar un algoritmo de aprendizaje en el conjunto de datos de entrenamiento para crear uno o más clasificadores entrenados, en donde cada clasificador entrenado clasifica una muestra de prueba en una o más de la pluralidad de clases.

A modo de ejemplo, un clasificador entrenado puede usar un algoritmo de aprendizaje seleccionado del grupo que consiste en: un bosque aleatorio, una red neuronal, una máquina de vectores de soporte y un clasificador lineal. Cada una de la pluralidad de clases diferentes puede seleccionarse del grupo que consiste en sano, cáncer de mama, cáncer de colon,

cáncer de pulmón, cáncer de páncreas, cáncer de próstata, cáncer de ovario, melanoma y cáncer hepático.

Un clasificador entrenado puede aplicarse a un método de clasificación de una muestra de un sujeto. Este método de clasificación puede comprender: (a) proporcionar un modelo multiparamétrico representativo de las moléculas de ADN libres de células de una muestra de prueba del sujeto; y (b) clasificar la muestra de prueba usando un clasificador entrenado. Una vez clasificada la muestra de prueba en una o más clases, se puede realizar una intervención terapéutica en el sujeto basada en la clasificación de la muestra.

En algunas modalidades, los conjuntos de entrenamiento se proporcionan a una unidad de aprendizaje automático, tal como una red neuronal o una máquina de vectores de soporte. Usando el conjunto de entrenamiento, la unidad de aprendizaje automático puede generar un modelo para clasificar la muestra según una respuesta de tratamiento a una o más invenciones terapéuticas. También se denomina "llamada". El modelo desarrollado puede emplear información de cualquier parte de un vector de prueba.

En general, el aprendizaje automático puede usarse para reducir un conjunto de datos generados a partir de todas las combinaciones (muestra primaria/analitos/prueba) a un conjunto predictivo óptimo de atributos, por ejemplo, que satisfagan criterios especificados. En varios ejemplos, puede aplicarse el aprendizaje estadístico y/o el análisis de regresión. En un paradigma de validación cruzada, pueden aplicarse a los datos modelos de simples a complejos y de pequeños a grandes, con diversos supuestos de modelización. De simple a complejo incluye consideraciones de linealidad a no linealidad y de representaciones no jerárquicas a jerárquicas de los atributos. Los modelos de pequeño a gran tamaño incluyen consideraciones sobre el tamaño del espacio del vector base sobre el que se proyectan los datos, así como el número de interacciones entre atributos que se incluyen en el proceso de modelización.

Pueden usarse técnicas de aprendizaje automático para valorar las modalidades de pruebas comerciales más óptimas en cuanto a coste/rendimiento/alcance comercial, como se definen en la pregunta inicial. Se puede realizar una comprobación del umbral: Si el método aplicado a un conjunto de datos de retención que no se usó en la validación cruzada supera las restricciones inicializadas, posteriormente se bloquea el ensayo y se inicia la producción. Por ejemplo, un umbral para el rendimiento del ensayo puede incluir una precisión mínima deseada, un valor predictivo positivo (VPP, por sus siglas en inglés), un valor predictivo negativo (VPN, por sus siglas en inglés), una sensibilidad clínica, una especificidad clínica, un

área bajo la curva (ABC) o una combinación de los mismos. Por ejemplo, una exactitud mínima deseada, VPP, VPN, sensibilidad clínica, especificidad clínica, o una combinación de las mismas puede ser de al menos aproximadamente 50 %, al menos aproximadamente 55 %, al menos aproximadamente 60 %, al menos aproximadamente 65 %, al menos aproximadamente 70 %, al menos aproximadamente 75 %, al menos aproximadamente 80 %, al menos aproximadamente 81 %, al menos aproximadamente 82 %, al menos aproximadamente 83 %, al menos aproximadamente 84 %, al menos aproximadamente 85 %, al menos aproximadamente 86 %, al menos aproximadamente 87 %, al menos aproximadamente 88 %, al menos aproximadamente 89 %, al menos aproximadamente 90 %, al menos aproximadamente 91 %, al menos aproximadamente 92 %, al menos aproximadamente 93 %, al menos aproximadamente 94 %, al menos aproximadamente 95 %, al menos aproximadamente 96 %, al menos aproximadamente 97 %, al menos aproximadamente 98 %, o al menos aproximadamente 99 %. Como otro ejemplo, un ABC mínima deseada puede ser de al menos aproximadamente 0,50, al menos aproximadamente 0,55, al menos aproximadamente 0,60, al menos aproximadamente 0,65, al menos aproximadamente 0,70, al menos aproximadamente 0,75, al menos aproximadamente 0,80, al menos aproximadamente 0,81, al menos aproximadamente 0,82, al menos aproximadamente 0,83, al menos aproximadamente 0,84, al menos aproximadamente 0,85, al menos aproximadamente 0,86, al menos aproximadamente 0,87, al menos aproximadamente 0,88, al menos aproximadamente 0,89, al menos aproximadamente 0,90, al menos aproximadamente 0,91, al menos aproximadamente 0,92, al menos aproximadamente 0,93, al menos aproximadamente 0,94, al menos aproximadamente 0,95, al menos aproximadamente 0,96, al menos aproximadamente 0,97, al menos aproximadamente 0,98, o al menos aproximadamente 0,99. Puede seleccionarse un subconjunto de ensayos a partir de un conjunto de ensayos a realizar en una muestra dada, basado en el coste total de realizar el subconjunto de ensayos, sujeto al umbral de rendimiento del ensayo, tal como la precisión mínima deseada, el valor predictivo positivo (VPP), el valor predictivo negativo (VPN), la sensibilidad clínica, la especificidad clínica, el área bajo la curva (ABC) y una combinación de los mismos. Si no se cumplen los umbrales, posteriormente el procedimiento de ingeniería del ensayo puede volver a la configuración de restricciones para una posible relajación o al laboratorio húmedo para cambiar los parámetros en los que se adquirieron los datos. Dada la pregunta clínica, las limitaciones biológicas, el presupuesto, las máquinas de laboratorio, etc., pueden restringir el problema.

En ciertas modalidades, el proceso informático de una técnica de aprendizaje automático puede incluir métodos de estadística, matemáticas, biología o cualquier

combinación de los mismos. En varios ejemplos, cualquiera de los métodos de procesamiento informático puede incluir un método de reducción de dimensiones, regresión logística, reducción de dimensiones, análisis de componentes principales, autocodificadores, descomposición de valores singulares, bases de Fourier, descomposición de valores singulares, ondículas, análisis discriminante, máquina de vectores de soporte, métodos basados en árboles, bosque aleatorio, árbol de impulso de gradiente, regresión logística, factorización de matrices, agrupación de redes, pruebas estadísticas y red neuronal.

En ciertas modalidades, el proceso informático de una técnica de aprendizaje automático puede incluir regresión logística, regresión lineal múltiple (MLR), reducción de dimensiones, regresión por mínimos cuadrados parciales (PLS), regresión por componentes principales, autocodificadores, autocodificadores variacionales, descomposición de valores singulares, bases de Fourier, ondículas, análisis discriminante, máquina de vectores soporte, árbol de decisión, árboles de clasificación y regresión (CART), métodos basados en árboles, bosque aleatorio, árbol de refuerzo de gradiente, regresión logística, factorización de matrices, escalado multidimensional (MDS), métodos de reducción de la dimensionalidad, incrustación estocástica de vecinos distribuida en t (t-SNE), perceptrón multicapa (MLP), agrupación en red, neuro-difuso, redes neuronales (superficiales y profundas), redes neuronales artificiales, coeficiente de correlación producto-momento de Pearson, coeficiente de correlación de rango de Spearman, coeficiente de correlación de rango tau de Kendall, o cualquier combinación de los mismos. En algunos ejemplos, el método de procesamiento informático es un método de aprendizaje automático supervisado que incluye, por ejemplo, un método de regresión, una máquina de vectores de soporte, un método basado en árboles y una red neuronal. En algunos ejemplos, el método de procesamiento informático es un método de aprendizaje automático no supervisado que incluye, por ejemplo, agrupación, grafos, análisis de componentes principales y factorización de matrices.

Para el aprendizaje supervisado, las muestras de entrenamiento (por ejemplo, en miles) pueden incluir datos medidos (por ejemplo, de varios analitos) y etiquetas conocidas, que pueden determinarse mediante otros procesos que requieren mucho tiempo, tal como la imagenología del sujeto y el análisis por un profesional capacitado. Las etiquetas de ejemplo pueden incluir la clasificación de un sujeto, por ejemplo, la clasificación discreta de si un sujeto tiene cáncer o no o clasificaciones continuas que proporcionan una probabilidad (por ejemplo, un riesgo o una puntuación) de un valor discreto. Un módulo de aprendizaje puede optimizar los parámetros de un modelo tal que se alcance una métrica de calidad (por ejemplo, la precisión de la predicción de una etiqueta conocida) con uno o más criterios especificados.

Determinar una métrica de calidad puede aplicarse a cualquier función arbitraria, incluido el conjunto de todas las funciones de riesgo, pérdida, utilidad y decisión. Se puede usar un gradiente junto con un paso de aprendizaje (por ejemplo, una medida de cuánto deben actualizarse los parámetros del modelo para un determinado paso temporal del proceso de optimización).

Como se ha descrito anteriormente, los ejemplos pueden usarse para diversos fines. Por ejemplo, se puede recolectar plasma (u otra muestra) de sujetos sintomáticos con una condición (por ejemplo, que se sabe que tienen la condición) y de sujetos sanos. Los datos genéticos (por ejemplo, ADNlc) pueden adquirirse analizados para obtener una variedad de atributos diferentes, que pueden incluir atributos basados en un análisis del genoma completo. Estos atributos pueden formar un espacio de atributos que se busca, estira, rota, traslada y transforma lineal o no linealmente para generar un modelo de aprendizaje automático exacto, que puede diferenciar entre sujetos sanos y sujetos con la condición (por ejemplo, identificar el estado de enfermedad o no enfermedad de un sujeto). Los resultados derivados de estos datos y del modelo (que pueden incluir probabilidades de la condición, etapas (niveles) de la condición u otros valores), pueden usarse para generar otro modelo que puede usarse para recomendar otros procedimientos, por ejemplo, recomendar una biopsia o seguir monitorizando la condición del sujeto.

En algunas modalidades, el ADN de una población de varios individuos puede analizarse por un conjunto de matrices multiplexados. Los datos de cada matriz multiplexada pueden autonormalizarse usando la información contenida en esa matriz específica. Este algoritmo de normalización puede ajustar las variaciones de intensidad nominal observadas en los canales de dos colores, las diferencias de fondo entre los canales y la posible diafonía entre los tintes. El comportamiento de cada posición de base puede modelarse posteriormente mediante un algoritmo de agrupación que incorpora varias heurísticas biológicas sobre perfiles de fragmentación. En los casos en los que se observan pocos fragmentos de ADNlc (por ejemplo, debido a una baja frecuencia de alelos menores), las ubicaciones y formas de las secuencias que faltan pueden estimarse usando redes neuronales. Dependiendo de los perfiles y del porcentaje de identidad de secuencia, se puede elaborar una puntuación estadística (una Puntuación de Entrenamiento). Una puntuación tal como GenCall Score está diseñado para imitar las evaluaciones realizadas por los sistemas visuales y cognitivos de un experto humano. Además, se ha evolucionado usando los datos de genotipado de las hebras superior e inferior. Este puntaje puede combinarse con varios términos de penalización (por ejemplo, baja intensidad, apareamiento erróneo entre los fragmentos de ADNlc existentes y

los previstos) para conformar la Puntuación de Entrenamiento. La Puntuación de Entrenamiento se guarda para ser usada por el algoritmo de llamada.

5 Para llamar a una respuesta terapéutica, un algoritmo de llamada puede tomar la información genética y las respuestas al tratamiento de una pluralidad de individuos que tienen una enfermedad o condición. Los datos pueden normalizarse primero (usando el mismo procedimiento que para el algoritmo de agrupación). La operación de llamada (clasificación) puede realizarse usando, por ejemplo, un modelo bayesiano. La puntuación de cada llamada puede ser el producto de una Puntuación de Entrenamiento y una puntuación de ajuste de datos a modelo. Después de puntuar todas las respuestas de tratamiento, la aplicación puede
10 calcular una puntuación compuesta.

En algunas modalidades, un conjunto de datos de entrenamiento comprende datos clínicos seleccionados del grupo que consiste en la etapa del cáncer, el tipo de procedimiento quirúrgico, la edad, la gradación del tumor, la profundidad de la infiltración tumoral, la aparición de complicaciones postoperatorias y la presencia de invasión venosa. En algunas
15 modalidades, el conjunto de datos de entrenamiento se preprocesa, lo que incluye la transformación de los datos proporcionados en probabilidades condicionales de clase.

Otra modalidad usa técnicas de aprendizaje automático para entrenar un clasificador estadístico, específicamente una máquina de vectores de soporte, para cada categoría de etapa del cáncer basada en ocurrencias de palabras en un corpus de informes histológicos para cada paciente. Los nuevos informes pueden clasificarse posteriormente en función de la
20 etapa más probable, lo que facilita la recopilación y el análisis de datos sobre la estadificación de la población.

En algunas modalidades, se selecciona un algoritmo de aprendizaje automático del grupo que consiste en: un algoritmo de aprendizaje supervisado o no supervisado
25 seleccionado entre máquina de vectores soporte, bosque aleatorio, análisis del vecino más próximo, regresión lineal, árbol de decisión binario, análisis discriminante, clasificador logístico y análisis de conglomerados.

En general, un sistema puede comprender un generador de informes para informar sobre los resultados de las pruebas del cáncer y las opciones de tratamiento. El sistema
30 generador de informes puede ser un sistema central de procesamiento de datos configurado para establecer comunicaciones directamente con: un sitio remoto de datos o laboratorio, un consultorio médico/proveedor de servicios de salud (profesional tratante) y/o un

paciente/sujeto a través de enlaces de comunicación. El laboratorio puede ser un laboratorio médico, un laboratorio de diagnóstico, un centro médico, una consulta médica, un dispositivo de pruebas en el punto de atención o cualquier otro sitio de datos remoto capaz de generar información clínica del sujeto. La información clínica del sujeto incluye, pero no se limita a,

5 datos de pruebas de laboratorio, radiografías, exámenes y diagnósticos. El proveedor de asistencia sanitaria o consulta 26 incluye a los proveedores de servicios médicos, tales como médicos, enfermeros, auxiliares sanitarios a domicilio, técnicos y asistentes médicos, y la consulta es cualquier centro de atención médica dotado de personal sanitario. En ciertos casos, el proveedor de atención sanitaria o la consulta también son sitios de datos remotos.

10 En una modalidad para el tratamiento del cáncer, el sujeto puede estar afectado de cáncer, entre otros.

Otra información clínica para un sujeto con cáncer incluye los resultados de pruebas de laboratorio, imagenología o procedimientos médicos dirigidos al cáncer específico que una persona con conocimientos ordinarios en la materia puede identificar fácilmente. La lista de

15 fuentes apropiadas de información clínica para el cáncer incluye, pero no se limita a: TAC, resonancia magnética, ecografía, gammagrafía ósea, PET, prueba de médula ósea, radiografía de bario, endoscopia, linfangiograma, IVU (urograma intravenoso) o IVP (pielograma IV), punción lumbar, cistoscopia, pruebas inmunológicas (detección de anticuerpos antimalignina) y pruebas de marcadores de cáncer; (por sus siglas en inglés,

20 respectivamente).

La información clínica del sujeto puede obtenerse del laboratorio de forma manual o automática. Para simplificar el sistema, la información se obtiene automáticamente a rangos de tiempo predeterminados o regulares. Un rango de tiempo regular se refiere a un rango de tiempo en el que la recolección de los datos de laboratorio se lleva a cabo automáticamente

25 por los métodos y sistemas descritos en la presente basado en una medida de tiempo tal como horas, días, semanas, meses, años, etc. En una modalidad de la invención, la recolección de datos y el proceso se llevan a cabo al menos una vez al día. En una modalidad, la transferencia y recolección de datos se lleva a cabo una vez al mes, quincenalmente, o una vez a la semana, o una vez cada dos días. Alternativamente, la recuperación de información

30 puede realizarse a rangos de tiempo predeterminados pero no regulares. Por ejemplo, un primer paso de recuperación puede ocurrir después de una semana y un segundo paso de recuperación puede ocurrir después de un mes. La transferencia y recopilación de datos puede personalizarse en función de la naturaleza del trastorno que se esté tratando y de la frecuencia de las pruebas y exámenes médicos necesarios de los sujetos.

En ciertas modalidades, se genera un informe genético a partir de una muestra del sujeto, por ejemplo, ADNlc. Los polinucleótidos de una muestra pueden secuenciarse, por ejemplo, mediante secuenciación del genoma completo, secuenciación NGS (por sus siglas en inglés), produciendo una pluralidad de lecturas de secuencias. En algunas modalidades, la información genética comprende variables que definen la organización genómica de las células cancerosas o la organización genómica de células cancerosas diseminadas individuales. En algunas modalidades, la información genética comprende datos de secuencia o abundancia de uno o más loci genéticos en ADN libre de células de los individuos.

Se procesa la información genética del ADNlc (72). También pueden identificarse variantes genéticas. Las variantes genéticas incluyen variantes de secuencia, variantes de número de copias y variantes de modificación de nucleótidos. Una variante de secuencia es una variación en una secuencia genética de nucleótidos. Una variante del número de copias es una desviación del tipo silvestre en el número de copias de una porción de un genoma. Las variantes genéticas incluyen, por ejemplo, variaciones de nucleótido único (SNP, por sus siglas en inglés), inserciones, deleciones, inversiones, transversiones, translocaciones, fusiones génicas, fusiones cromosómicas, truncamientos génicos, variaciones en el número de copias (por ejemplo, aneuploidía, aneuploidía parcial, poliploidía, amplificación génica), cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos y cambios anormales en la metilación de los ácidos nucleicos. El proceso determina posteriormente la frecuencia de las variantes genéticas en la muestra que contiene el material genético. Dado que este proceso es ruidoso, el proceso separa la información del ruido (73). La sensibilidad de la detección de variantes genéticas puede aumentarse incrementando la profundidad de lectura de los polinucleótidos (por ejemplo, secuenciando a una mayor profundidad de lectura en una muestra de un sujeto en dos o más puntos temporales).

Para aumentar la confianza del diagnóstico, se puede tomar una pluralidad de mediciones. O, alternativamente, usar mediciones en una pluralidad de puntos temporales (por ejemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10 o más puntos temporales) para determinar si el cáncer está avanzando, en remisión o estabilizado. La confianza diagnóstica puede usarse para identificar estados de enfermedad. Por ejemplo, los polinucleótidos libres de células tomados de un sujeto pueden incluir polinucleótidos derivados de células normales, así como polinucleótidos derivados de células enfermas, tales como células cancerosas. Los polinucleótidos de células cancerosas pueden presentar variantes genéticas, tales como mutaciones somáticas celulares y variantes en el número de copias. Cuando se secuencian

polinucleótidos libres de células de una muestra de un sujeto, se pueden producir perfiles de fragmentación del ADNlc como se describe en la sección de ejemplos a continuación.

5 Pueden detectarse numerosos tipos de cáncer usando los métodos y sistemas descritos en la presente. Las células cancerosas, como la mayoría de las células, pueden caracterizarse por una tasa de recambio, en la que las células viejas mueren y son sustituidas por células nuevas. Por lo general, las células muertas, en contacto con la vasculatura de un sujeto determinado, pueden liberar ADN o fragmentos de ADN en el torrente sanguíneo. Lo mismo ocurre con las células cancerosas en las distintas etapas de la enfermedad. Las células cancerosas también pueden caracterizarse, dependiendo de la etapa de la enfermedad, por
10 diversas aberraciones genéticas tales como variaciones en el número de copias, así como mutaciones. Este fenómeno puede usarse para detectar la presencia o ausencia de individuos con cáncer usando los métodos y sistemas descritos en la presente.

En la detección precoz de cánceres, cualquiera de los sistemas o métodos en la presente descritos, incluida la detección de mutaciones o la detección de variaciones en el
15 número de copias, puede usarse para detectar cánceres. Estos sistemas y métodos pueden usarse para detectar cualquier número de aberraciones genéticas que puedan causar o resultar en cánceres. Pueden incluir, pero no se limitan a, perfiles de fragmentación del ADNlc, mutaciones, mutaciones, indels, variaciones del número de copias, transversiones, translocaciones, inversiones, deleciones, aneuploidías, aneuploidías parciales, poliploidías,
20 inestabilidad cromosómica, alteraciones de la estructura cromosómica, fusiones génicas, fusiones cromosómicas, truncamientos génicos, amplificación génica, duplicaciones génicas, lesiones cromosómicas, lesiones del ADN, cambios anormales en las modificaciones químicas de los ácidos nucleicos, cambios anormales en los patrones epigenéticos, cambios anormales en la metilación de los ácidos nucleicos infección y cáncer.

25 Adicionalmente, los sistemas y métodos descritos en la presente también pueden usarse para ayudar a caracterizar ciertos tipos de cáncer. Los datos genéticos producidos a partir del sistema y los métodos de esta divulgación pueden permitir a los profesionales ayudar a caracterizar mejor una forma específica de cáncer. A menudo, los cánceres son heterogéneos tanto en su composición como en su etapa. Los datos del perfil genético pueden
30 permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes para el diagnóstico o el tratamiento de ese subtipo específico. Esta información también puede proporcionar al sujeto o al médico pistas sobre el pronóstico de un tipo específico de cáncer.

Los sistemas y métodos proporcionados en la presente pueden ser usados para

monitorizar cánceres ya conocidos, u otras enfermedades en un sujeto particular. Esto puede permitir al sujeto o al profesional adaptar las opciones de tratamiento en función de la evolución de la enfermedad. En este ejemplo, los sistemas y métodos descritos en la presente pueden usarse para construir perfiles genéticos de fragmentación del ADNlc de un sujeto particular del curso de la enfermedad. En algunos casos, los cánceres pueden progresar, volviéndose más agresivos y genéticamente inestables. En otros ejemplos, los cánceres pueden permanecer benignos, inactivos o latentes. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad.

Además, los sistemas y métodos descritos en la presente pueden ser útiles para determinar la eficacia de una opción de tratamiento particular. En un ejemplo, ciertas opciones de tratamiento pueden correlacionarse con los perfiles genéticos de fragmentación del ADNlc de los cánceres a lo largo del tiempo. Esta correlación puede ser útil a la hora de seleccionar una terapia. Adicionalmente, si se observa que un cáncer está en remisión después del tratamiento, los sistemas y métodos descritos en la presente pueden ser útiles para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

Además, los métodos de la divulgación pueden usarse para caracterizar la heterogeneidad de una condición anormal en un sujeto, comprendiendo el método generar un perfil de fragmentación del ADNlc de polinucleótidos extracelulares en el sujeto, en donde el perfil de fragmentación del ADNlc comprende una pluralidad de datos resultantes de los análisis de variación y mutación del perfil. En algunos casos, incluido el cáncer pero sin limitarse al mismo, una enfermedad puede ser heterogénea. Las células enfermas pueden no ser idénticas. En el ejemplo del cáncer, se sabe que algunos tumores comprenden diferentes tipos de células tumorales, algunas células en diferentes etapas del cáncer. En otros ejemplos, la heterogeneidad puede comprender múltiples focos de enfermedad. De nuevo, en el ejemplo del cáncer, puede haber múltiples focos tumorales, tal vez cuando uno o más focos son el resultado de metástasis que se han extendido desde un sitio primario (también conocidas como metástasis a distancia).

Los métodos de esta divulgación pueden usarse para generar un perfil, huella digital o conjunto de datos que sea una suma de información genética derivada de diferentes células en una enfermedad heterogénea. Este conjunto de datos puede incluir análisis de variación del número de copias y de mutaciones, solos o combinados.

Además, estos informes se presentan y consultan electrónicamente mediante Internet. El análisis de los datos se realiza en un lugar distinto al del sujeto. El informe se genera y se

transmite al lugar donde se encuentra el sujeto. Mediante una computadora con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral.

La información anotada puede ser usada por un proveedor de atención médica para seleccionar otras opciones de tratamiento con fármacos y/o proporcionar información sobre
5 opciones de tratamiento con fármacos a una compañía de seguros. El método puede incluir la anotación de las opciones de tratamiento con fármacos para una condición en, por ejemplo, las NCCN Clinical Practice Guidelines in Oncology™ (por su nombre en inglés) o las guías de práctica clínica de la American Society of Clinical Oncology (ASCO, por sus siglas en inglés).

Se generan informes, mapeando las posiciones del genoma y la variación del perfil de
10 fragmentación del ADNlc para el sujeto con cáncer. Estos informes, en comparación con otros perfiles de sujetos con resultados conocidos, pueden indicar que un cáncer en particular es agresivo y resistente al tratamiento. Se monitoriza al sujeto durante un periodo y se le vuelven a realizar pruebas. Si al final del periodo, el perfil de variación de la fragmentación del ADNlc no varía, esto puede indicar que el tratamiento actual no está funcionando. Se realiza una
15 comparación con los perfiles de fragmentación del ADNlc de otros sujetos. Por ejemplo, si se determina que un cambio en la variación de la fragmentación del ADNlc indica que el cáncer está avanzando, posteriormente el régimen de tratamiento original prescrito ya no está tratando el cáncer y se prescribe un nuevo tratamiento.

En ciertas modalidades, el sistema recibe información genética de un secuenciador de
20 ADN. El proceso determina posteriormente alteraciones específicas de la fragmentación del ADNlc y sus cantidades. Estos informes se presentan y consultan electrónicamente mediante Internet. El análisis de los datos se realiza en un lugar distinto al del sujeto. El informe se genera y se transmite al lugar donde se encuentra el sujeto. Mediante una computadora con acceso a Internet, el sujeto accede a los informes que reflejan su carga tumoral.

Aunque la información temporal puede usarse para potenciar la información de los
25 perfiles de fragmentación del ADNlc, pueden aplicarse otros métodos de consenso. En otras modalidades, la comparación histórica puede usarse junto con otros perfiles de fragmentación del ADNlc consensuados. Los perfiles de fragmentación del ADNlc de consenso pueden normalizarse con respecto a las muestras de control. Las medidas del mapeo de las moléculas
30 con las secuencias de referencia también pueden compararse a lo largo de un genoma para identificar las áreas del genoma en las que los perfiles de fragmentación del ADNlc varían o permanecen invariables. Los métodos de consenso incluyen, por ejemplo, métodos lineales o no lineales de construcción de perfiles de fragmentación del ADNlc de consenso (tales como

la votación, el promedio, la detección estadística, máxima a posteriori o máxima verosimilitud, la programación dinámica, los métodos bayesianos, de Markov oculto o de máquina de vectores de soporte, etc.) derivados de la teoría de la comunicación digital, la teoría de la información o la bioinformática. Una vez determinada la cobertura de lectura de secuencia, se aplica un algoritmo de modelado estocástico para convertir la cobertura de lectura de secuencia de ácido nucleico normalizada para cada región de ventana en los estados discretos de número de copias. En algunos casos, este algoritmo puede comprender uno o más de los siguientes elementos: Modelo de Markov oculto, programación dinámica, máquina de vectores soporte, red bayesiana, decodificación trellis, decodificación Viterbi, maximización de expectativas, metodologías de filtrado Kalman y redes neuronales.

Las redes neuronales artificiales (RNA, por sus siglas en inglés) imitan las redes de “neuronas” basadas en la estructura neuronal del cerebro. Procesan los registros de uno en uno, o por lotes, y “aprenden” comparando su clasificación del registro (que, al principio, es en gran medida arbitraria) con la clasificación real conocida del registro. En las redes MLP, los errores de la clasificación inicial del primer registro se devuelven a la red y se usan para modificar el algoritmo de la red la segunda vez, y así sucesivamente durante muchas iteraciones. Las redes neuronales usan un proceso de aprendizaje iterativo en el que los casos de datos (filas) se presentan a la red de uno en uno, y los pesos asociados a los valores de entrada se ajustan cada vez.

Una vez presentados todos los casos, a menudo el proceso vuelve a empezar. Durante esta fase de aprendizaje, la red aprende ajustando los pesos para poder predecir la etiqueta de clase correcta de las muestras de entrada. El aprendizaje en redes neuronales también se denomina “aprendizaje conexionista”, debido a las conexiones entre las unidades. Entre las ventajas de las redes neuronales figuran su gran tolerancia a los datos ruidosos y su capacidad para clasificar patrones para los que no han sido entrenadas. Un algoritmo de red neuronal es el de retropropagación, tal como el de Levenberg-Marquadt. Una vez estructurada una red para una aplicación particular, esa red está lista para ser entrenada. Para iniciar este proceso, los pesos iniciales se eligen al azar. Posteriormente comienza la formación o el aprendizaje.

La red procesa los registros de los datos de entrenamiento de uno en uno, usando los pesos y las funciones de las capas ocultas, y posteriormente compara las salidas resultantes con las salidas deseadas. Posteriormente, los errores se propagan de vuelta por el sistema, haciendo que éste ajuste los pesos para aplicarlos al siguiente registro que se vaya a

procesar. Este proceso se repite una y otra vez, ya que los pesos se ajustan continuamente. Durante el entrenamiento de una red, el mismo conjunto de datos se procesa muchas veces, ya que los pesos de conexión se refinan continuamente.

5 En una modalidad, el paso de entrenamiento de la unidad de aprendizaje automático sobre el conjunto de datos de entrenamiento puede generar uno o más modelos de clasificación para aplicar a una muestra de prueba. Estos modelos de clasificación pueden aplicarse a una muestra de prueba para predecir la respuesta de un sujeto a una intervención terapéutica.

10 Comparar la cobertura de la secuencia con una muestra de control o secuencia de referencia puede ayudar a la normalización entre ventanas. En esta modalidad, el ADN libre de células se extrae y aísla de un fluido corporal fácilmente accesible tal como la sangre. Por ejemplo, el ADN libre de células puede extraerse usando una variedad de métodos conocidos en la materia, incluyendo pero no limitado a la precipitación con isopropanol y/o la purificación basada en sílice. El ADN libre de células puede extraerse de cualquier número de
15 sujetos, tales como sujetos sin cáncer, sujetos con riesgo de cáncer o sujetos que se sabe que tienen cáncer (por ejemplo, por otros medios).

Después del paso de aislamiento/extracción, puede realizarse cualquiera de las diferentes operaciones de secuenciación en la muestra de polinucleótidos libre de células. Las muestras pueden procesarse antes de la secuenciación con uno o más reactivos (por ejemplo, enzimas, identificadores únicos (por ejemplo, códigos de barras), sondas, etc.). En algunos
20 casos, si la muestra se procesa con un identificador único tal como un código de barras, las muestras o fragmentos de muestras pueden etiquetarse individualmente o en subgrupos con el identificador único. La muestra etiquetada puede posteriormente usarse en una aplicación de cadena abajo tal como una reacción de secuenciación por la cual las moléculas individuales
25 pueden rastrearse hasta las moléculas progenitoras.

Los polinucleótidos libres de células pueden etiquetarse o rastrearse para permitir subsecuentemente la identificación y el origen del polinucleótido en particular. La asignación de un identificador (por ejemplo, un código de barras) a individuos o subgrupos de polinucleótidos puede permitir asignar una identidad única a secuencias individuales o
30 fragmentos de secuencias. Esto puede permitir la adquisición de datos de muestras individuales y no se limita a promedios de muestras. En algunos ejemplos, los ácidos nucleicos u otras moléculas derivadas de una hebra pueden compartir una etiqueta o identificador común y, por lo tanto, pueden identificarse posteriormente como derivados de esa hebra.

Similarmente, todos los fragmentos de una hebra única de ácido nucleico pueden marcarse con el mismo identificador o etiqueta, permitiendo así la identificación subsecuente de los fragmentos de la hebra madre. En otros casos, los productos de la expresión génica (por ejemplo, el ARNm) pueden etiquetarse para cuantificar la expresión, para lo cual puede
5 contarse el código de barras, o el código de barras en combinación con la secuencia a la que está unido. En otros casos, los sistemas y métodos pueden usarse como control de amplificación por PCR (por sus siglas en inglés). En tales casos, múltiples productos de amplificación de una reacción de PCR pueden ser etiquetados con la misma etiqueta o identificador. Si los productos se secuencian posteriormente y muestran diferencias de
10 secuencia, las diferencias entre productos con el mismo identificador pueden atribuirse entonces a un error de PCR. Adicionalmente, pueden identificarse secuencias individuales basadas en características de los datos de secuencia de la propia lectura. Por ejemplo, la detección de datos de secuencia únicos en las porciones inicial (inicio) y final (parada) de lecturas de secuenciación individuales puede usarse, sola o en combinación, con la longitud,
15 o el número de pares de bases de cada lectura de secuencia única para asignar identidades únicas a moléculas individuales. Los fragmentos de una hebra única de ácido nucleico, a los que se ha asignado una identidad única, pueden por lo mismo permitir la identificación subsecuente de fragmentos de la hebra parental. Esto puede usarse junto con el estrangulamiento del material genético inicial para limitar la diversidad.

20 En general, los métodos y sistemas proporcionados en la presente son útiles para la preparación de secuencias de polinucleótidos libres de células para una reacción de secuenciación de aplicación descendente. A menudo, un método de secuenciación es la secuenciación de nueva generación (NGS), la secuenciación clásica de Sanger, la secuenciación de bisulfito de genoma completo (WGSB, por sus siglas en inglés), la
25 secuenciación de ARN pequeño, la secuenciación de genoma completo de baja cobertura (IcWGS, por sus siglas en inglés), etc.

Como se usa en la presente, el término “secuenciación” se refiere a cualquiera de las diversas tecnologías usadas para determinar la secuencia de una biomolécula, por ejemplo, un ácido nucleico tal como ADN o ARN. Los métodos de secuenciación ejemplares incluyen,
30 pero no se limitan a, secuenciación dirigida, secuenciación en tiempo real de molécula única, secuenciación de exones, secuenciación basada en microscopía electrónica, secuenciación de panel, secuenciación mediada por transistor, secuenciación directa, secuenciación aleatoria de escopeta, secuenciación por terminación dideoxi de Sanger, secuenciación del genoma completo, secuenciación por hibridación, pirosecuenciación, electroforesis capilar,

electroforesis en gel, secuenciación duplexada, secuenciación por ciclos, secuenciación por extensión monobase, secuenciación en fase sólida, secuenciación de alto rendimiento, secuenciación de firmas masivamente paralela, PCR en emulsión, coamplificación a baja temperatura de desnaturalización-PCR (COLD-PCR, por su nombre en inglés), PCR multiplexada, secuenciación por tinte reversible terminador, secuenciación de terminadores pareados, secuenciación a corto plazo, secuenciación por exonucleasa, secuenciación por ligadura, secuenciación de lectura corta, secuenciación de molécula única, secuenciación por síntesis, secuenciación en tiempo real, secuenciación por terminador inverso, secuenciación nanopore, secuenciación 454, secuenciación Solexa Genome Analyzer, secuenciación SOLiD™, secuenciación MS-PET, y una combinación de las mismas. En algunas modalidades, la secuenciación puede llevarse a cabo por un analizador de genes tal como, por ejemplo, los analizadores de genes disponibles comercialmente en Illumina o Applied Biosystems. En algunas modalidades, el método de secuenciación puede ser secuenciación paralela masiva, es decir, secuenciación simultánea (o en rápida sucesión) de al menos 100, 1000, 10.000, 100.000, 1 millón, 10 millones, 100 millones o 1000 millones de moléculas de polinucleótidos.

Después de la secuenciación, se asigna una puntuación de calidad a las lecturas. Una puntuación de calidad puede ser una representación de lecturas que indica si esas lecturas pueden ser útiles en análisis subsecuentes basados en un umbral. En algunos casos, algunas lecturas no son de suficiente calidad o longitud para realizar el subsecuente paso de mapeo. Las lecturas de secuenciación con una puntuación de calidad de al menos 90 %, 95 %, 99 %, 99,9 %, 99,99 % o 99,999 % pueden filtrarse del conjunto de datos. En otros casos, las lecturas de secuenciación a las que se haya asignado una puntuación de calidad de al menos 90 %, 95 %, 99 %, 99,9 %, 99,99 % o 99,999 % pueden filtrarse del conjunto de datos. En el paso 306, las lecturas de fragmentos genómicos que cumplen un umbral de puntaje de calidad especificado se mapean a un genoma de referencia, o a una secuencia de referencia que se sabe que no contiene mutaciones. Después del alineamiento de mapeo, a las lecturas de secuencias se les asigna una puntuación de mapeo. Una puntuación de mapeo puede ser una representación o lecturas mapeadas de vuelta a la secuencia de referencia indicando si cada posición es o no mapeable de forma única. En algunos casos, las lecturas pueden ser secuencias no relacionadas con el análisis de mutaciones. Por ejemplo, algunas lecturas de secuencias pueden proceder de polinucleótidos contaminantes. Las lecturas de secuenciación con una puntuación de mapeo de al menos 90 %, 95 %, 99 %, 99,9 %, 99,99 % o 99,999 % pueden filtrarse del conjunto de datos. En otros casos, las lecturas de secuenciación a las que se haya asignado una puntuación de mapeo menor a 90 %, 95 %, 99

%, 99,9 %, 99,99 % o 99,999 % pueden filtrarse del conjunto de datos. Para cada base mapeable, las bases que no alcanzan el umbral mínimo de mapeabilidad, o las bases de baja calidad, pueden sustituirse por las bases correspondientes tal y como se encuentran en la secuencia de referencia.

5 Pueden detectarse numerosos tipos de cáncer usando los métodos y sistemas descritos en la presente. Las células cancerosas, como la mayoría de las células, pueden caracterizarse por una tasa de recambio, en la que las células viejas mueren y son sustituidas por células nuevas. Por lo general, las células muertas, en contacto con la vasculatura de un sujeto determinado, pueden liberar ADN o fragmentos de ADN en el torrente sanguíneo. Lo mismo ocurre con las células cancerosas en las distintas etapas de la enfermedad. Las células cancerosas también pueden caracterizarse, dependiendo de la etapa de la enfermedad, por diversas aberraciones genéticas tales como variaciones en el número de copias, así como mutaciones. Este fenómeno puede usarse para detectar la presencia o ausencia de individuos con cáncer usando los métodos y sistemas descritos en la presente.

15 Los tipos y el número de cánceres que pueden detectarse pueden incluir, pero no se limitan a, cánceres de sangre, cánceres de cerebro, cánceres de pulmón, cánceres de piel, cánceres de nariz, cánceres de garganta, cánceres hepáticos, cánceres de hueso, linfomas, cánceres de páncreas, cánceres de piel, cánceres de intestino, cánceres de recto, cánceres de tiroides, cánceres de vejiga, cánceres de riñón, cánceres de boca, cánceres de estómago, tumores en estado sólido, tumores heterogéneos, tumores homogéneos y similares.

Adicionalmente, los sistemas y métodos descritos en la presente también pueden usarse para ayudar a caracterizar ciertos tipos de cáncer. Los datos genéticos producidos a partir del sistema y los métodos de esta divulgación pueden permitir a los profesionales ayudar a caracterizar mejor una forma específica de cáncer. A menudo, los cánceres son heterogéneos tanto en su composición como en su etapa. Los datos del perfil genético pueden permitir la caracterización de subtipos específicos de cáncer que pueden ser importantes para el diagnóstico o el tratamiento de ese subtipo específico. Esta información también puede proporcionar al sujeto o al médico pistas sobre el pronóstico de un tipo específico de cáncer.

Los sistemas y métodos proporcionados en la presente pueden ser usados para monitorizar cánceres ya conocidos, u otras enfermedades en un sujeto particular. Esto puede permitir al sujeto o al profesional adaptar las opciones de tratamiento en función de la evolución de la enfermedad. En este ejemplo, los sistemas y métodos descritos en la presente pueden usarse para construir perfiles genéticos de un sujeto particular del curso de la

enfermedad. En algunos casos, los cánceres pueden progresar, volviéndose más agresivos y genéticamente inestables. En otros ejemplos, los cánceres pueden permanecer benignos, inactivos o latentes. El sistema y los métodos de esta divulgación pueden ser útiles para determinar la progresión de la enfermedad.

5 Además, los sistemas y métodos descritos en la presente pueden ser útiles para determinar la eficacia de una opción de tratamiento particular. En un ejemplo, las opciones de tratamiento exitosas pueden en realidad aumentar la cantidad de variación del número de copias o mutaciones detectadas en la sangre del sujeto si el tratamiento tiene éxito, ya que más cánceres pueden morir y desprender ADN. En otros ejemplos, esto puede no ocurrir. En
10 otro ejemplo, quizá ciertas opciones de tratamiento puedan correlacionarse con los perfiles genéticos de los cánceres a lo largo del tiempo. Esta correlación puede ser útil a la hora de seleccionar una terapia. Adicionalmente, si se observa que un cáncer está en remisión después del tratamiento, los sistemas y métodos descritos en la presente pueden ser útiles para monitorizar la enfermedad residual o la recurrencia de la enfermedad.

15 Los datos se envían a través de una conexión directa o de Internet a una computadora para su proceso. Los aspectos de procesamiento de datos del sistema pueden implementarse en circuitos electrónicos digitales o en hardware informático, firmware, software o en combinaciones de ellos. El aparato de procesamiento de datos de la invención puede implementarse en un producto de programa de computadora tangiblemente incorporado en
20 un dispositivo de almacenamiento legible por máquina para su ejecución por un procesador programable; y los pasos del método de procesamiento de datos de la invención pueden ser realizados por un procesador programable que ejecuta un programa de instrucciones para realizar funciones de la invención operando sobre datos de entrada y generando salida. Los aspectos de procesamiento de datos de la invención pueden implementarse ventajosamente
25 en uno o más programas informáticos que son ejecutables en un sistema programable que incluye al menos un procesador programable acoplado para recibir datos e instrucciones desde y para transmitir datos e instrucciones a un sistema de almacenamiento de datos, al menos un dispositivo de entrada y al menos un dispositivo de salida. Cada programa informático puede implementarse en un lenguaje de programación de alto nivel, procedimental
30 u orientado a objetos, o en lenguaje ensamblador o de máquina, si se desea; y, en cualquier caso, el lenguaje puede ser un lenguaje compilado o interpretado. Los procesadores adecuados incluyen, a modo de ejemplo, microprocesadores de uso general y especial. Generalmente, un procesador recibirá instrucciones y datos de una memoria de solamente lectura y/o de una memoria de acceso aleatorio. Los dispositivos de almacenamiento

adecuados para incorporar de forma tangible instrucciones y datos de programas informáticos incluyen todas las formas de memoria no volátil, incluyendo a modo de ejemplo dispositivos de memoria semiconductores, tales como EPROM, EEPROM y dispositivos de memoria flash; discos magnéticos tales como discos duros internos y discos extraíbles; discos magneto-
5 ópticos; y discos CD-ROM (por sus siglas en inglés, respectivamente). Todo lo anterior puede complementarse con ASIC (circuitos integrados para aplicaciones específicas) o incorporarse a ellos.

Para facilitar la interacción con el usuario, los métodos pueden implementarse usando un sistema informático que tiene un dispositivo de visualización, tal como un monitor o una
10 pantalla LCD (pantalla de cristal líquido, por sus siglas en inglés), para mostrar información al usuario, y dispositivos de entrada mediante los cuales el usuario pueda proporcionar información al sistema informático, tal como un teclado, un dispositivo señalador bidimensional, tal como un ratón o un rueda de desplazamiento, o un dispositivo señalador tridimensional, tal como un guante de datos o un ratón giroscópico. El sistema informático
15 puede programarse para proporcionar una interfase gráfica de usuario a través de la cual los programas informáticos interactúan con los usuarios. El sistema informático puede programarse para proporcionar una interfase de visualización tridimensional de realidad virtual.

EJEMPLOS

20 Ejemplo 1: Cohortes clínicas y análisis genómicos del ADNlc

Se examinaron muestras de plasma de 501 individuos, incluidos 75 individuos con HCC y 426 sin cáncer. Entre los individuos sin cáncer, 133 presentaban condiciones que aumentaban el riesgo de HCC, incluida la cirrosis por todas las causas o la hepatitis viral sin cirrosis. Se recolectaron prospectivamente muestras de sangre de pacientes con HCC en
25 diversas etapas del cáncer y de individuos de alto riesgo en el Hospital Johns Hopkins, mientras que las muestras restantes se identificaron a través de esfuerzos de examinación en otros hospitales de EE.UU. o de la UE (cohorte EE.UU./UE) (Tabla 1). Aislamos de 0,5 a 5 mL de plasma de cada uno de estos individuos, generamos bibliotecas genómicas y secuenciamos los fragmentos de ADNlc usando secuenciación del genoma completo de baja
30 cobertura (~2,6x de cobertura) con un promedio de 49 millones de lecturas emparejadas de alta calidad por muestra que comprenden 9 Gb de datos de secuencia (24, 25). Además de la cohorte EE.UU./UE, examinamos como cohorte de validación los datos de la secuencia del genoma completo de 223 pacientes de Hong Kong, incluidos pacientes con HCC resecable

en etapa temprana (n = 90, etapa A = 85, B = 5), HBV (n = 66) y cirrosis relacionada con el HBV (n = 35), así como individuos sanos sin enfermedad hepática (n = 32) (cohorte Hong Kong) (Tabla 1) (15, 28).

5 Perfiles de fragmentación del ADNlc en el genoma completo reportados por la estructura subyacente de la cromatina

10 Evaluamos el fragmentoma y generamos perfiles de fragmentación a lo largo del genoma en 473 regiones no solapadas de 5 MB, cada región compuesta por ~80,000 fragmentos, y abarcando aproximadamente 2,4 GB del genoma usando el enfoque DELFI (24). Los perfiles de fragmentación eran coherentes entre los individuos sin cáncer, pero muy variables entre los pacientes con HCC (figura 1A). Los perfiles de los pacientes con cirrosis eran más parecidos a los de las personas sin cáncer y sin cirrosis que a los de los pacientes con HCC (figura 1A). Asimismo, los pacientes con hepatitis viral presentaban perfiles de fragmentación casi idénticos a los de las personas sin cáncer y sin enfermedad hepática (figura 1A).

15 Para examinar los orígenes de los patrones de fragmentación del ADNlc, comparamos los perfiles del fragmentoma del genoma completo con los compartimentos abiertos (A) y cerrados (B) de la captura de la conformación cromosómica (Hi-C) por secuenciación de alto rendimiento. Encontramos que los patrones del ADNlc de individuos sanos estaban altamente correlacionados con los de las células linfoblastoides (figura 1B). El análisis de los perfiles del
20 ADNlc de 10 pacientes de HCC con niveles elevados de ADNlc reveló que su fragmentoma reflejaba dos componentes, uno parecido al perfil de los individuos sin cáncer, así como un componente separado del ADNlc que tenía una gran similitud con los compartimentos A/B estimados previamente de los cánceres hepáticos (figura 1B) (29). Adicionalmente, cuando se estimaron estos dos componentes, los perfiles del ADNlc del componente hepático previsto
25 tenían una gran similitud con los compartimentos A/B del genoma completo del cáncer hepático, mientras que los perfiles de los pacientes con HCC tenían una similitud intermedia con el cáncer hepático (figura 1B, figura 1C). Por el contrario, los perfiles de los individuos sin cáncer se acercaban más a los compartimentos A/B de las células linfoblastoides (figura 1B, figura 1C). Estos análisis sugirieron que los fragmentomas del ADNlc de individuos con HCC
30 representan una mezcla de perfiles del ADNlc de compartimentos de cromatina de células procedentes de sangre periférica, así como de células procedentes de cáncer hepático.

Factores de transcripción específicos de enfermedades inferidos a partir de la fragmentación del ADNlc en el genoma completo

Dado que la organización de la cromatina refleja los programas transcripcionales celulares subyacentes (30-32), examinamos si las características de fragmentación del ADNlc podrían reflejar cambios derivados de la alteración de la unión al ADN de los factores de transcripción (TF) en el cáncer hepático. Para identificar los sitios de unión al ADN de todos los TF conocidos, analizamos 5620 experimentos CHIP-seq de la base de datos ReMap 2020 (33). Para cada TF, calculamos la cobertura agregada de ADNlc en todos los sitios de unión identificados (4K - 490K por muestra) en comparación con la cobertura genómica adyacente global, produciendo una única métrica para cada TF en cada muestra. Comparamos estos TF en pacientes con y sin HCC para identificar aquellos TF con las mayores y menores diferencias en la cobertura de sitios de unión en el genoma completo en ADNlc (figura 2A, figura 2B). Los análisis de enriquecimiento de conjuntos de genes usando la base de datos DisGeNET de asociaciones gen-enfermedad revelaron que se predijo que las diferencias en las coberturas de unión a TF del ADNlc entre el HCC y los individuos sin cáncer estaban relacionadas con el cáncer hepático y otros cánceres (figura 2C, figura 2D). Adicionalmente, los TF individuales con mayor puntuación representaron a aquellos con relevancia biológica conocida para la organización de la cromatina y el cáncer hepático (Tabla 2). Entre ellos se incluyen miembros del complejo de la proteína activadora 1 (AP1), como los genes JUN, JUND, ATF2 y ATF7, que integran señales extracelulares (34) y se han vinculado a la tumorigénesis hepática (35, 36), el miembro de la familia del dominio del factor potenciador de transcripción 4 (TEAD4), que ha demostrado tener funciones oncogénicas en el HCC (37, 38), La proteína 2 de unión a poli(C) (PCBP2), un corregulador transcripcional que, cuando se sobreexpresa, se asocia a un peor pronóstico en pacientes con HCC (39), la inhibitina 2 (PHB), que promueve la progresión del HCC (40), y el dominio de interacción rico en AT 3A (ARID3A), un factor de transcripción oncogénico que, cuando se regula positivamente, promueve la malignidad del cáncer hepático (41); (por sus siglas en inglés, respectivamente). Un análisis similar de los datos de fragmentación de ADNlc de nuestro reciente estudio de pacientes en el ensayo de diagnóstico de cáncer de pulmón LUCAS (por sus siglas en inglés) (24) reveló un enriquecimiento de las diferencias de cobertura en los sitios de unión de TF relacionados con el cáncer de pulmón (figura 2C, figura 2E). En conjunto, estas observaciones sugieren que los cambios en la fragmentación del ADNlc en pacientes con cáncer hepático y otros tipos de cáncer resultan en la multitud de perfiles transcripcionales alterados presentes en las células cancerosas.

Los fragmentomas del ADNlc revelan cambios genómicos en el HCC

Dado que el fragmentoma del ADNlc puede comprender cambios relacionados con

alteraciones genómicas a gran escala liberadas de las células cancerosas (24, 25), también examinamos las ganancias y pérdidas cromosómicas en la circulación de estos pacientes. Además de los perfiles de fragmentación del genoma completo resultantes de los cambios en la cromatina y en los factores de transcripción observados en pacientes con cáncer hepático (figura 3A), nuestros análisis revelaron una representación alterada de los brazos cromosómicos que coinciden con los que comúnmente se ganan o se pierden en el cáncer hepático, como se informó en estudios genómicos anteriores a gran escala del TCGA sobre el HCC (n = 372) (figura 3B). Estos incluían un aumento de la representación en el ADNlc de 1q, 7p, 7q, 8q y una disminución de los niveles de 4q, 8p, 9p, 13q y 21q, todos ellos conocidos por ganarse o perderse, respectivamente, en el HCC (42, 43). Es importante destacar que estas alteraciones se observaron en los pacientes con HCC pero no en los individuos sin cáncer, aunque tuvieran cirrosis o enfermedad hepática crónica (figura 3B).

Modelo DELFI para la detección del HCC

Dada la conexión directa entre los cambios genómicos y cromatínicos en el cáncer hepático y la fragmentación del ADNlc, usamos un enfoque de aprendizaje automático para determinar si los cambios en los fragmentomas del ADNlc podían distinguir a los pacientes con HCC de los que no tenían cáncer. Anteriormente usamos este enfoque para desarrollar un clasificador robusto para la detección del cáncer de pulmón que fue validado externamente en una población independiente (24). Determinamos el rendimiento de este clasificador en la cohorte EE.UU./UE repitiendo la validación cruzada cinco veces, generando una puntuación para cada individuo que es un promedio de diez repeticiones de validación cruzada (puntuación DELFI). El modelo resultante incluía una combinación de características de fragmentación regional y a gran escala que resultaban óptimas para identificar a los individuos con cáncer hepático (figura 5). Estos atributos comprenden la mayoría de los cambios cromosómicos, cromatínicos y locales informativos identificados anteriormente, y representan más de 90 % de la varianza de los perfiles de fragmentación entre las muestras.

Dado que las características clínicas pueden afectar a los biomarcadores tumorales, se investigó si las medidas de disfunción hepática o los parámetros demográficos tales como la edad, el sexo, la raza o el peso se asociaban a las puntuaciones DELFI en individuos sin cáncer en los que se disponía de esta información.

No se observó ninguna asociación entre las puntuaciones DELFI y la edad ($R = 0,18$, $p = 0,08$, correlación de Spearman) (figura 6A) y ninguna diferencia en las puntuaciones DELFI entre hombres y mujeres ($p = 0,58$, prueba de Wilcoxon) (figura 6B). Se ha demostrado que

los asiáticos y los afroamericanos tienen una mayor incidencia de cáncer hepático diagnosticado en etapas más avanzadas (44), y observamos pequeñas diferencias en las puntuaciones de fragmentación entre individuos de alto riesgo sin cáncer en estos u otros grupos raciales o étnicos, aunque estos análisis están limitados por la falta de información sobre covariables clínicas en algunos de estos casos ($p = 0,037$ en pacientes con hepatitis viral, y $p = 0,026$ en pacientes cirróticos, prueba de Kruskal-Wallis) (figura 7). Entre los individuos con cirrosis, observamos una correlación entre el grado de enfermedad hepática medido por la puntuación Child-Pugh y las puntuaciones DELFI ($R = 0,58$ $p = 8,6e-5$, correlación de Spearman) (figura 8). El aumento del índice de masa corporal (IMC), un factor de riesgo de NAFLD y cáncer hepático, no se asoció a cambios en las puntuaciones DELFI en pacientes con hepatitis viral ($R = 0,027$, $p = 0,85$, correlación de Spearman); sin embargo, un menor IMC en pacientes cirróticos se asoció a mayores puntuaciones DELFI, quizá debido a la caquexia en pacientes con cirrosis grave ($R = -0,23$, $p = 0,043$, correlación de Spearman) (figura 9).

A continuación examinamos la relación entre las puntuaciones DELFI y la presencia y etapa del cáncer hepático en una población con alto riesgo de padecerlo. Las puntuaciones DELFI de 133 individuos libres de cáncer fueron bajas, con medianas de puntuaciones DELFI de 0,078 o 0,080 para aquellos con hepatitis viral o cirrosis, respectivamente. Por el contrario, los 75 pacientes con HCC presentaron puntuaciones DELFI medias significativamente más altas en todas las etapas del BCLC, incluyendo la etapa 0 = 0,46, la etapa A = 0,61, la etapa B = 0,83 y la etapa C = 0,92 ($p < 0,01$ para las etapas 0, A, B o C, prueba de suma de rangos de Wilcoxon, figura 4A). Una curva de característica del operador receptor (ROC) del enfoque DELFI para identificar a los pacientes con HCC reveló un área bajo la curva (ABC) de 0,90 (CI de 95 % = 0,86-0,94) entre los individuos de alto riesgo. El rendimiento siguió siendo sólido para los HCC en etapa temprana, con un ABC de 0,9 y 0,81 para las etapas 0 y A de BCLC. Los HCC en etapa avanzada (BCLC C) se detectaron casi perfectamente entre los individuos analizados (ABC > 0,97) (figura 4C).

Para ampliar estos análisis a los individuos con bajo riesgo de desarrollar cáncer hepático, examinamos la capacidad de un modelo DELFI para distinguir entre los individuos con cáncer y los de una población general ($n = 293$) sin hepatitis viral ni cirrosis. En esta cohorte más amplia en la que se podían incluir atributos adicionales en el entrenamiento de validación cruzada, usamos los atributos del modelo anterior e incluimos además la cobertura del ADNlc en los sitios de unión de factores de transcripción derivados de CHIP-seq de líneas celulares hepáticas disponibles en la base de datos ReMAP para crear un modelo DELFI para

una población general. Este enfoque tuvo un alto rendimiento para la detección del cáncer ($ABC = 0,98$) entre estos individuos. Evaluamos el rendimiento de este modelo con una especificidad de 99 %, un umbral apropiado para una población de riesgo medio (25), y observamos una sensibilidad global de 80 % en este contexto (figura 4B), con una sensibilidad mayor a 65 % en todas las etapas. Usar un modelo que no incorpore sitios de unión a factores de transcripción dio lugar a un rendimiento ligeramente menor, y hubo una alta correlación entre las puntuaciones ordenados por rango usando nuestros modelos DELFI para poblaciones de alto riesgo y de examinación ($R = .64$, $p < 1e-15$) (figura 10 y figura 11).

Para examinar la relación entre los perfiles de fragmentación y la progresión del cáncer hepático, se valoró si el tamaño, el número y las características de las lesiones del cáncer hepático, así como la etiología de la neoplasia, estaban relacionados con los perfiles de fragmentación aberrantes, cuando se disponía de esta información. Se observó que el tamaño del tumor y el número de lesiones se correlacionaban positivamente con las puntuaciones DELFI ($R = 0,42$ y $.31$, $p = 0,00026$ y $p = 0,0064$ respectivamente, correlación de Spearman) (figura 12), en consonancia con la idea de que el perfil de fragmentación estaba relacionado con la carga tumoral global.

Entre los pacientes con cáncer hepático en etapas resecables (0, A y B), la etiología del cáncer, incluida la hepatitis viral, o la cirrosis debida al alcohol, la NAFLD o de origen idiopático, arrojaron puntuaciones DELFI similares ($p = 0,43$, prueba de Kruskal-Wallis) (figura 13). Estas observaciones sugieren que los perfiles de fragmentación eran el resultado de procesos continuos relacionados con el ADNlc tumoral y no se veían afectados por eventos tempranos de la tumorigénesis.

Para examinar el impacto en el mundo real de este método en el contexto de la detección del HCC, comparamos el rendimiento del fragmentoma DELFI con la medición de examinación actual de los niveles de proteína alfa-fetoproteína (AFP). Los niveles de AFP se elevaron por encima del umbral de examinación recomendado de 20 ng/mL en 39 de 75 (52 %) individuos con cáncer, en consonancia con informes anteriores (45). Entre los individuos que tenían niveles de AFP menores a 20 ng/mL y que no habían sido detectados por este método, el DELFI detectó 30 de 36 (83 %). Usando mediciones de AFP se habrían detectado 8/24 (33 %) pacientes en etapa 0/A, 17/30 (57 %) de pacientes en etapa B y 14/21 (66 %) pacientes en etapa C (figura 14). Por el contrario, el enfoque DELFI detectó 19/24 (79 %) pacientes en etapa 0/A, 25/30 (83 %) pacientes en etapa B y 20/21 (95 %) pacientes en etapa C. En general, los análisis de fragmentación del ADNlc en el genoma completo tuvieron un

mejor rendimiento en comparación con la detección del HCC por AFP, y la combinación de DELFI y AFP puede proporcionar una mejora en la detección con respecto al enfoque DELFI solo, ya que observamos que tenían una sensibilidad combinada de 92 % con una especificidad combinada de 80 %.

5 **Validación externa del modelo DELFI en población de Asia oriental con HCC**

Además de nuestro análisis de validación cruzada de la cohorte EE.UU./UE, probamos el modelo DELFI fijo en los 223 pacientes de la cohorte Hong Kong. Entre ellos había pacientes con un HCC en fase inicial en gran medida resecable ($n = 90$, etapa A = 85, B = 5) y 101 con cirrosis o infección por HBV. Estas muestras se secuenciaron previamente usando un secuenciador diferente (HiSeq 2000 contra Novaseq; 76 pb contra 100 pb de longitud de lectura), una preparación de biblioteca diferente y un mayor número de ciclos de PCR (14 contra 4 ciclos), pero observamos patrones similares en el genoma completo a los de nuestros análisis anteriores (figura 15). Los perfiles de fragmentación de los pacientes con hepatitis viral y cirrosis, así como de los individuos sanos, presentaban perfiles muy coherentes en el genoma completo, mientras que los de los pacientes con HCC eran variables y desordenados (figura 16). Adicionalmente, los cambios cromosómicos observados en el plasma de la cohorte Hong Kong fueron similares a los de la cohorte inicial de EE.UU./UE, así como a los cánceres de TCGA (figura 17). En general, en esta cohorte de validación, el modelo DELFI distinguió a los pacientes con HCC con un ABC de 0,97 de aquellos con enfermedad de alto riesgo (figura 4D). Estas observaciones sugieren que las características subyacentes de la fragmentación del ADNlc fueron similares en esta cohorte, y que DELFI es un método robusto para detectar el HCC y es generalizable en diferentes poblaciones de alto riesgo.

Simulación del funcionamiento de DELFI a escala poblacional

Para evaluar cómo funcionaría nuestro enfoque para la vigilancia y detección en pacientes con alto riesgo de cáncer hepático, evaluamos el modelo DELFI en una población teórica de 100.000 individuos de alto riesgo usando simulaciones Monte Carlo. Dada la importancia de la detección de cánceres en etapas tempranas, centramos nuestro modelo en la detección de la enfermedad en etapa 0/A. Comparamos el enfoque DELFI con el estándar de atención actual, ecografía y AFP concurrentes, y modelamos la incertidumbre de la sensibilidad y especificidad de estas modalidades de vigilancia en esta población teórica mediante distribuciones de probabilidad centradas en estimaciones empíricas de nuestra cohorte o de informes anteriores ((11), ver Métodos). A pesar de las recomendaciones de vigilancia, la adherencia a la vigilancia del HCC en EE.UU. es baja, y las estimaciones más

generosas sugieren una adherencia de 39 % (46), lo que resulta en un promedio de 40.042 individuos sometidos a prueba en esta población teórica (CI de 95 %, 21.320-61.890). Dado que las pruebas sanguíneas ofrecen una accesibilidad y un cumplimiento elevados, con tasas de adherencia de 80-90 % notificadas para biomarcadores sanguíneos (47, 48), asumimos de forma conservadora que un promedio de 75 % (CI de 95 %, 60-90 %) de esta población se sometería a las pruebas usando el enfoque DELFI. Dado que la prevalencia de cirrosis, hepatitis viral y la coexistencia de estas comorbilidades con el HCC podría variar según la región, usamos una distribución de probabilidad a priori para reflejar nuestra incertidumbre sobre la composición de estas enfermedades y las posibles diferencias regionales. Las simulaciones de Monte Carlo a partir de estas distribuciones de probabilidad (Métodos) revelaron que la ecografía con AFP detectó un promedio de 2233 (95 % de CI, 1088-3699) individuos con cáncer hepático (figura 18). Usando DELFI, detectaríamos un promedio de 2794 casos adicionales de cáncer hepático, o un aumento de 2,46 veces (95 % de CI, 1,25-4,57 veces) en comparación con la ecografía con AFP sola (figura 18). El enfoque DELFI no solamente mejoraría sustancialmente la detección del cáncer hepático, sino que se esperaría que disminuyera la tasa de falsos negativos (FNR, por sus siglas en inglés), o fracción de cánceres no detectados en la prueba, de 38 % para la ecografía con AFP (95 % de CI, 25 %-51,5 %) a 24 % para DELFI (95 % de CI, 9 %-42,6 %). Adicionalmente, cabría esperar que el valor predictivo negativo de la prueba (VPN) aumentara de 95,7 % para la ecografía con AFP (95 % de CI, 93,8 %-97,3 %) a 97,1 % para DELFI (95 % de CI, 94,8 %-99,0 %). Estos análisis sugieren un beneficio significativo para toda la población al usar una prueba de detección precoz de alta especificidad basada en sangre como herramienta para la detección del cáncer hepático.

Discusión

En general, en este estudio demostramos el uso de los atributos del fragmentoma del ADNc del genoma completo para detectar el HCC con alta sensibilidad y especificidad. Asimismo, demostramos que los perfiles de fragmentación captan características genómicas y cromatínicas, incluidas alteraciones que se sabe que son importantes en el HCC. Nuestro enfoque del fragmentoma del ADNc tiene un rendimiento sólido en la detección del HCC, incluida la enfermedad en etapas muy tempranas, independientemente de la etiología de la enfermedad. Hasta donde sabemos, este es el primer análisis de fragmentación del genoma completo que ha sido validado de forma independiente en una población de alto riesgo separada, con un rendimiento estable y robusto en diferentes grupos raciales y étnicos de EE.UU. y Hong Kong.

Nuestros resultados también revelaron que pueden obtenerse firmas de factores de transcripción específicas de la enfermedad mediante el análisis de perfiles de fragmentación del ADNlc del genoma completo. Aunque se han realizado análisis de este tipo usando factores de transcripción específicos para distinguir los carcinomas microcíticos de pulmón de los no microcíticos (24), este estudio sugiere que los análisis de la regulación transcripcional específica de la enfermedad usando la fragmentación del ADNlc en el genoma completo pueden mejorar la detección e identificación del tejido de origen en pacientes con cáncer. Con un número suficiente de pacientes, los perfiles transcripcionales del ADNlc podrían mejorar aún más los algoritmos de aprendizaje automático para detectar el HCC y otros tipos de cáncer.

El HCC es único en comparación con otros cánceres sólidos en el sentido de que existe una amplia población de alto riesgo bien definida con un riesgo anual promedio de 3-4 % de desarrollar HCC (49) a la que se recomienda someterse a un examen rutinario del cáncer cada seis meses. Desgraciadamente, las pruebas disponibles en la actualidad tienen una utilidad diagnóstica limitada, sobre todo en las primeras etapas de la enfermedad (13). En nuestro estudio, la AFP tuvo una sensibilidad de 52 % en la detección de HCC, en consonancia con el rendimiento conocido de este biomarcador (13). La vigilancia basada en ecografía también tiene limitaciones técnicas por su dependencia del operador y su menor sensibilidad en pacientes obesos y con cirrosis (50). Y lo que es más importante, la ecografía tiene un bajo cumplimiento de las guías establecidas, menos de 20 % en todo el mundo (10, 11), en comparación con un cumplimiento mucho mayor de las pruebas de sangre para otras condiciones (47). A pesar de estas dificultades, la examinación del HCC proporciona un beneficio de supervivencia global en pacientes con HBV (51) y cirrosis (10), lo que pone de manifiesto la gran necesidad de mejorar las pruebas de examinación actuales. El alto rendimiento de los análisis del fragmentoma del ADNlc en la detección del HCC, junto con sus características eficientes en cuanto a costes, permitiría que DELFI fuera una prueba de examinación accesible para el HCC y aumentara las tasas de examinación más allá de los pésimos niveles actuales. Un aspecto interesante del análisis del ADNlc específico para el HCC es que el trasplante es el tratamiento más curativo para el HCC en etapa temprana a intermedia, y la vigilancia del HCC en pacientes postrasplante con un enfoque de biopsia líquida podría tener la doble función de rastrear la recurrencia y el rechazo, ya que los estudios del ADNlc en pacientes postrasplante han demostrado ser prometedores (52).

Aunque este estudio representa una mejora potencial de los métodos de examinación actuales, presenta algunas limitaciones. Por ejemplo, este estudio incluyó una muestra

relativamente pequeña de individuos con HCC. Aunque la cohorte de validación independiente se realizó con diferencias preanalíticas en los métodos de laboratorio y secuenciación, el hecho de que el enfoque DELFI funcionara bien en esta población sugiere que el método podrá usarse en última instancia en un rango de diferentes laboratorios de diagnóstico. Serán necesarios estudios de validación más amplios antes de que este enfoque pueda usarse clínicamente. No obstante, la observación de que los análisis no invasivos del fragmentoma del ADNlc, escalables y rentables, pueden detectar a los pacientes con cáncer hepático puede brindar la oportunidad de examinar a las poblaciones de alto riesgo y a la población general en todo el mundo.

10 **Materiales y métodos**

Población de estudio

Para la cohorte EE.UU./UE, se recolectaron prospectivamente muestras de 208 pacientes, incluidos 75 con HCC y 133 pacientes de alto riesgo sin HCC, como parte del registro de biomarcadores de HCC y del estudio Sida vinculado a la experiencia intravenosa (ALIVE, por sus siglas en inglés) en la Facultad de Medicina de la Universidad Johns Hopkins, de acuerdo con protocolos aprobados por la Junta de Revisión Institucional de Johns Hopkins. El HCC se definió mediante un examen histológico o las características de imagenología apropiadas definidas por las directrices aceptadas. La etapa tumoral se determinó mediante el sistema de estadificación del cáncer hepático de la Clínica de Barcelona (BCLC, por sus siglas en inglés). Los datos clínicos detallados se extrajeron de la historia clínica electrónica.

Los pacientes de alto riesgo se definieron como individuos con cirrosis de cualquier etiología y/o individuos con hepatitis crónica B o hepatitis C a los que las directrices de la sociedad de expertos recomendaban la examinación sistemático del HCC (50). Además, se incluyeron 38 pacientes con hepatitis B o cirrosis recolectados retrospectivamente por BioIVT (Westbury, NY). Los niveles de AFP fueron cuantificados por los centros asociados en sus laboratorios clínicos usando pruebas de AFP aprobadas por la Administración de Alimentos y Fármacos.

La cohorte EE.UU./UE también incluyó muestras de 293 individuos sin cáncer que se analizaron previamente (24), procedentes originalmente de dos cohortes de ensayos clínicos de examinación del cáncer colorrectal en Dinamarca (Endoscopia III) y los Países Bajos (COCOS, ID del registro de ensayos de los Países Bajos NTR182946). El protocolo del Proyecto Endoscopia III fue aprobado por el Comité Ético Regional y la Agencia Danesa de

Protección de Datos, y para el ensayo COCOS se obtuvo la aprobación ética del Consejo Holandés de Salud. Los criterios de inclusión tanto para la cohorte holandesa como para la danesa fueron cualquier individuo de 50-75 años elegible para la examinación de cáncer colorrectal. Todos los pacientes usados tenían una prueba FIT (por sus siglas en inglés) negativa o un resultado de colonoscopia negativo.

Para la cohorte Hong Kong, todos los sujetos reclutados dieron su consentimiento informado por escrito, y el estudio fue aprobado por el Comité Ético de Investigación Clínica de la Universidad China Conjunta de Hong Kong y el Conjunto de Nuevos Territorios del Este (15, 28).

10 **Recolección y conservación de muestras**

La recolección de muestras se realizó del siguiente modo: Se recolectó sangre venosa periférica en un tubo de K2-EDTA (por sus siglas en inglés) y dos tubos de gel de suero. En las dos horas siguientes a la extracción de sangre, los tubos se centrifugaron a 2330 g a 4 °C por 10 minutos, el plasma se transfirió a tubos nuevos y las muestras se centrifugaron a 14.000 rpm (18.000 rcf) por 10 minutos a temperatura ambiente para precipitar cualquier resto celular. Después de la centrifugación, se tomaron alícuotas de plasma con EDTA y se almacenaron a -80 °C para los análisis del ADNlc.

Preparación de bibliotecas de secuenciación

El ADN circulante libre de células se aisló a partir de 2-4 mL de plasma usando el kit Qiagen QIAamp Circulating Nucleic Acids (Qiagen GmbH), se eluyó en 52 µL de agua libre de ARNasa que contenía 0,04 % de azida de sodio(Qiagen GmbH) y se almacenó en tubos LoBind (Eppendorf AG) a -20 °C. La concentración y la calidad del ADNlc se usaron para su valoración en el Bioanalyzer 2100 (Agilent Technologies).

Las bibliotecas de ADNlc para secuenciación de próxima generación (NGS) se prepararon para WGS usando 15 ng del ADNlc cuando estaban disponibles, o la cantidad purificada completa cuando era menor a 15 ng. En resumen, las bibliotecas genómicas se prepararon usando el kit NEBNext DNA Library Prep Kit for Illumina (New England Biolabs (NEB, por sus siglas en inglés)) con cuatro modificaciones principales respecto a las guías del fabricante: (i) los pasos de purificación de bibliotecas siguieron el enfoque en-perlas AMPure XP (Beckman Coulter) para minimizar la pérdida de muestras durante los pasos de elución y transferencia de tubos; (ii) los volúmenes de enzimas y amortiguadores de reparación de

extremos NEBNext, cola de A y ligación de adaptadores se ajustaron como proceda para adaptarse a la purificación en-perlas AMPure XP; (iii) se usaron adaptadores de índice dual Illumina en la reacción de ligación; y (iv) las bibliotecas de ADNlc se amplificaron con la polimerasa Phusion Hot Start. Todas las muestras se sometieron a una amplificación PCR de 4 ciclos después del paso de ligadura del ADN.

Secuenciación y alineamiento del genoma completo de baja cobertura

Las bibliotecas de genomas completos de pacientes con cáncer y de individuos libres de cáncer se prepararon como en (24), con la modificación de que se secuenciaron usando series de 100 pb (200 ciclos) en la plataforma Illumina Novaseq con una cobertura de 1-2x por genoma. Antes del alineamiento, se filtraron las secuencias adaptadoras de las lecturas usando el software fastp (53). Las lecturas de secuencias se alinearon con el genoma de referencia humano hg19 usando Bowtie2 (54) y las lecturas duplicadas se eliminaron usando Sambamba (55). Después del alineamiento, cada par alineado se convirtió en un rango genómico que representaba el fragmento de ADN secuenciado usando bedtools (por su nombre en inglés) (56). Solamente se retuvieron las lecturas con una puntuación MAPQ igual o mayor a 30. Los pares de lecturas se filtraron además si se solapaban con la lista negra de regiones excluidas de Duke (genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeMapability). Para capturar las diferencias epigenéticas a gran escala en la fragmentación a lo largo del genoma estimable a partir de WGS (por sus siglas en inglés) de baja cobertura, dividimos el genoma de referencia hg19 en paquetes de 5 Mb no solapados. Se excluyeron los paquetes con un contenido promedio de GC < 0,3 y una mapeabilidad promedio < 0,9, lo que dejó 473 paquetes que abarcaban aproximadamente 2,4 GB del genoma. Siguiendo a Mathios *et al.* (24), la corrección de GC se realizó de forma independiente para fragmentos cortos (= 150 pb) del ADNlc usando un panel externo de 20 individuos sin cáncer secuenciados en un NovaSeq para generar una distribución objetivo.

Los archivos Fastq de los pacientes de la cohorte Hong Kong se obtuvieron del Grupo de Investigación de Ácidos Nucleicos Circulantes de la Universidad China de Hong Kong (CUHK, por sus siglas en inglés), como se informó (15, Jiang, 2018 #1645) y se procesaron como se describió anteriormente y en Mathios *et. al.*, para generar las características DELFI. La corrección del GC se realizó normalizando a la distribución objetivo proporcionada en github.com/cancer-genomics/PlasmaToolsNovaseq.hg19, la misma distribución objetivo usada para la corrección del GC en la cohorte EE.UU./UE. El conjunto de validación consistió en bibliotecas construidas con 14 ciclos de PCR y secuenciadas en el HiSeq 2000. Estas

bibliotecas se normalizaron a la distribución objetivo de NovaSeq de 4 ciclos para facilitar las comparaciones entre estudios. Se excluyó una muestra del grupo cirrótico y otra del grupo HBV, ya que se identificó que tenían un diagnóstico de HCC.

Análisis de la estructura de la cromatina

5 Los compartimentos A/B para el tejido de cáncer hepático y las células linfoblastoides se obtuvieron de github.com/Jfortin1/TCGA_AB_Compartments, así como de github.com/Jfortin1/HiC_AB_Compartments, como se describe en (29). Se compararon las dos pistas de referencia para identificar los paquetes informativos de 100kb, definidos como paquetes en los que el dominio de cromatina difería entre las dos pistas de referencia o la
10 diferencia de magnitud en los valores propios que correspondía a una puntuación z mayor a 1,96 o menor a -1,96 ($p = 0,05$) en todas las diferencias de valores propios.

Se calculó la mediana del perfil de fragmentación de las 10 muestras de hígado con la mayor fracción tumoral estimada por ichorCNA y de 10 individuos sin cáncer seleccionados al azar. Esta información se usó para extraer una mediana estimada del componente hepático
15 en el plasma ponderada por la puntuación de icor de las muestras individuales de plasma.

Análisis de factores de transcripción en el genoma completo

Los picos de inmunoprecipitación de cromatina seguida de secuenciación (ChIP-Seq) de 5620 experimentos se descargaron de la base de datos ReMap 2020 (33). Este conjunto se filtró para los experimentos con más de 4000 picos, resultando en 4293 experimentos. Para
20 cada pico en los autosomas definimos el centro del pico como la posición 0.

Se calculó la media de las coberturas en cada posición (-3000 a +3000 con respecto al centro de cada pico) en todos los picos de cada muestra. Para las curvas ROC, la cobertura relativa se calculó para cada muestra como la cobertura media en una ventana de +/- 100 pb alrededor del centro de los sitios de unión dividida por la cobertura media en una ventana de
25 +/- 250 pb alrededor de 2750 pb cadena arriba y cadena abajo de los sitios de unión. La curva ROC se generó usando pROC 1.16.2 (58). Se clasificó el ABC de cada conjunto de picos. Cada factor de transcripción se emparejó con su ID NCBI, quedando 797 factores de transcripción únicos clasificados por el ABC. Esta lista ordenada fue la entrada para la función gseDGN del paquete DOSE en R. El resultado se ordenó por la puntuación de enriquecimiento
30 normalizado (NSE, por sus siglas en inglés).

Atributos de los fragmentos del genoma completo

Los atributos de fragmentación se calcularon como en Mathios et. al. (24). En resumen, se calculó la relación entre fragmentos cortos y largos en 473 paquetes de 5 MB no superpuestos en el genoma completo y se calcularon las puntuaciones z que representan las ganancias/pérdidas de brazos en los brazos cromosómicos autosómicos. Los componentes principales de las relaciones que representaban más de 90 % de la varianza y las puntuaciones z se usaron para entrenar modelos de aprendizaje automático.

Aprendizaje automático y análisis de validación cruzada

Se desarrollaron dos modelos de aprendizaje automático, uno para poblaciones de alto riesgo (un GBM, por sus siglas en inglés, usando los atributos de Mathios *et al*), y el segundo para poblaciones generales de bajo riesgo (una regresión logística penalizada con los atributos de Mathios *et al*, así como la cobertura de los sitios de unión de factores de transcripción). Estos modelos fueron entrenados en la cohorte EE.UU./UE en Caret con validación cruzada de 5 veces con 10 repeticiones, y las puntuaciones para cada muestra fueron calculadas por la media a través de repeticiones y evaluados usando ABC-ROC, como en Mathios et. al. (24). En el primer modelo se usaron los pacientes de alto riesgo sin cáncer y con HCC, mientras que en el segundo se usaron los individuos sin cáncer y sin patología hepática. El modelo bloqueado de alto riesgo entrenado en la cohorte EE.UU./UE se aplicó a la cohorte Hong Kong para generar predicciones de cáncer en un conjunto de validación externo.

Análisis TCGA

Los datos de número de copias de la cohorte de cáncer de HCC en TCGA (LIHC n = 372) se recuperaron usando el paquete RTCGA v1.16.0 y se analizaron para determinar la frecuencia de ganancias y pérdidas de número de copias en los 473 paquetes de 5 mb para esta cohorte (24). Se usó el umbral de alteración somática del número de copias (SCNA, por sus siglas en inglés) usado en Mathios *et al*. para llamar ganancias y pérdidas en las cohortes de HCC (24, 59).

Asociación de covariables clínicas con la puntuación DELFI

Las posibles asociaciones entre las covariables clínicas (para aquellos pacientes de los que se disponía de esta información) y la puntuación DELFI se evaluaron mediante el coeficiente de correlación de rangos de Spearman (variables continuas) y el análisis de varianza unidireccional de Kruskal-Wallis (variables categóricas).

Simulación

Se usaron simulaciones de Monte Carlo para comparar el enfoque DELFI con la ecografía y la AFP en una población de vigilancia teórica. Usamos intervalos de confianza de 95 % estimados de sensibilidad y especificidad para DELFI e intervalos de confianza de 95 % publicados para ecografía con AFP (13). Se usó el paquete R epiR para derivar distribuciones de probabilidad predictivas previas (distribuciones beta) a partir de estos intervalos de confianza ((60) paquete R versión 2.47, CRAN.R-project.org/package = Epi). Zhao *et al.* (2017) informaron de que la adherencia a la vigilancia de EE.UU. y AFP fue de 39 % (CI de 95 %: 21 %-65 %). Dado que en otras pruebas no invasivas basadas en la sangre se ha notificado una adherencia mayor a 75 % (47, 48), asumimos que la adherencia al DELFI sería de 60 % o mayor con una probabilidad de 0,975 o mayor. Usando estas estimaciones de confianza, epiR se usó para derivar distribuciones predictivas beta *a priori* para la adherencia. Se simularon probabilidades multinomiales para la prevalencia de hepatitis B, cirrosis, hepatitis B + HCC, cirrosis + HCC y hepatitis B + cirrosis + HCC a partir de un Dirichlet con parámetros 230, 680, 60, 23, 7, respectivamente. Para una única simulación de Monte Carlo para ecografía con prueba AFP, nosotros

- (i) muestreamos la probabilidad de adherencia (η) a partir de la distribución predictiva *a priori*,
- (ii) simulamos el número de 100.000 individuos (S) que participaron en la vigilancia ($S \sim \text{Binomial}(\eta, 100.000)$),
- (ii) muestreamos probabilidades de comorbilidades (Dirichlet (230, 680, 60, 23, 7)),
- (iv) calculamos la prevalencia del HCC (θ),
- (v) simulamos casos de HCC ($P \sim \text{Binomial}(\theta, S)$) y calcularon el número de individuos sin cáncer ($N = S - P$),
- (vi) muestreamos la sensibilidad (se) y la especificidad (esp) a partir de las correspondientes distribuciones predictivas *a priori*, y
- (vii) muestreamos los verdaderos positivos ($VP \sim \text{Binomial}(P, se)$) y los falsos positivos ($FP \sim \text{Binomial}(N, 1-esp)$).

- Dados VP y FP , calculamos el VNP como $(\text{verdaderos negativos})/(\text{verdaderos negativos} + \text{falsos negativos})$, en donde verdaderos negativos = $N - FP$ y falsos negativos = $P - VP$. Repetimos la simulación anterior 1000 veces, obteniendo una distribución de VP , FP y VNP . Usando parámetros de sensibilidad, especificidad y adherencia para el enfoque DELFI,
- 5 repetimos el mismo análisis de Monte Carlo para permitir comparaciones entre estas dos metodologías de vigilancia.

Tabla 1. Datos demográficos y clínicos del paciente

Características del paciente	Individuos cáncer n = 426	sin Pacientes cáncer n = 75	con Valor de p*
Edad			
Media	57,5	64,5	< 0,001
Rango	27-81	38-88	
Sexo			
Hombre	235	63	< 0,001
Mujer	191	12	
Enfermedad hepática			
Ninguna	293		< 0,001
Hepatitis B	26	1	
Hepatitis C	29	1	
Cirrosis	78	69	< 0,001
HCV	53	41	
HBV	2	4	
EtOH	13	12	
NAFLD	3	11	
Etapas de Child-Pugh			
A	20	43	< 0,001
B	10	38	
C	10	5	
Desconocida	38		
Etapas BCLC			
0		7	
A		17	
B		30	

C		21
Tratamiento previo		
Sí		28
No		47
Cohorte de validación (Hong Kong)[#]	133	90
Enfermedad hepática		
Ninguna	32	
Cirrosis (HBV)	35	90
HBV activo	66	
Etapas BCLC		
A		85
B		5

* Se calcularon los valores p para comparar los datos de los individuos con y sin cáncer hepático en relación con las siguientes variables: medias de edad por pruebas t de Student no apareadas de 2 colas, distribución sec, etiología de la cirrosis y etapa Child-Pugh por una prueba χ^2 .

[#] Los datos de la cohorte de validación se obtuvieron de Jiang *et al.* PNAS, 2015.

Tabla 2. Factores de transcripción con mayor puntuación en muestras de cohorte EE.UU./UE

Factor de transcripción	Nombre genético	ABC	Tipo de célula en el experimento CHIP-seq	Función genética*	Enlace a HCC
ZNF512	Proteína de dedos de zinc 512	0,836	K-562	Desconocida	No descrito
ZNF184	Proteína de dedos de zinc 184	0,826	K-562	Desconocida	No descrito
PCBP2	Proteína 2 de unión a poli(C)	0,791	Hep-G2	Corregulador traslacional	La sobreexpresión contribuye a un mal pronóstico y a potenciar el crecimiento celular en el HCC {Zhang, 2016}

JUN	Protooncogén Jun, subunidad del factor de transcripción AP-1	0,785	Hep-G2	TF	Promueve la tumorigénesis hepática relacionada con HBV {Trierweiler, 2016}
PHB2	Inhibitina 2	0,771	K-562	Corregulador traslacional	Funciones en la mitofagia del HCC {Xiang, 2021}
ATF2	Factor de transcripción activador 2	0,770	Hep-G2	TF, HAT	Interviene en la supresión de los tumores hepáticos
ATF7	Factor de transcripción activador 7	0,765	MCF-7	TF	Regula el crecimiento del cáncer hepático {Song, 2019}
TEAD4	Factor de transcripción 4 con dominio TEA	0,760	Hep-G2	TF	Papel oncogénico en el HCC {Coto-Llerena, 2021}
ARID3A	Dominio de interacción rico en AT 3A	0,756	Hep-G2	TF	Facilita la malignización del cáncer hepático {Shen, 2022}
JUND	Protooncogén Jun-D, subunidad del factor de transcripción AP-1	0,754	HT29_DMSO	TF	Señalización PARy y desarrollo de NAFLD {Hasenfusss, 2014}

*TF = factor de transcripción, HAT= histona acetil-transferasa.

Referencias

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36
5 cancers in 185 countries. *CA: a cancer journal for clinicians* **2021**;71(3):209-49.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA: a cancer journal for clinicians* **2021**;71(1):7-33.
3. Di Bisceglie AM. Hepatitis C and hepatocellular carcinoma. *Hepatology* **1997**;26(3 Suppl 1):34S-8S doi 10.1002/hep.510260706.
- 10 4. Pinyopornpanish K, Khoudari G, Saleh MA, Angkurawaranon C, Pinyopornpanish K, Mansoor E, *et al.* Hepatocellular carcinoma in nonalcoholic fatty liver disease with or without

cirrhosis: a population-based study. *BMC gastroenterology* **2021**;21(1):1-7.

5. Donato F, Tagger A, Gelatti U, Parrinello G, Boffetta P, Albertini A, *et al.* Alcohol and hepatocellular carcinoma: the effect of lifetime intake and hepatitis virus infections in men and women. *Am J Epidemiol* **2002**;155(4):323-31 doi 10.1093/aje/155.4.323.

5 6. Waly Raphael S, Yangde Z, Yuxiang C. Hepatocellular carcinoma: focus on different aspects of management. *ISRN Oncol* **2012**;2012:421673 doi 10.5402/2012/421673.

7. Asrani SK, Devarbhavi H, Eaton J, Kamath PS. Burden of liver diseases in the world. *J Hepatol* **2019**;70(1):151-71 doi 10.1016/j.jhep.2018.09.014.

8. Frenette CT, Isaacson AJ, Bargellini I, Saab S, Singal AG. A Practical Guideline for
10 Hepatocellular Carcinoma Screening in Patients at Risk. *Mayo Clin Proc Innov Qual Outcomes* **2019**;3(3):302-10 doi 10.1016/j.mayocpiqo.2019.04.005.

9. Kanwal F, Singal AG. Surveillance for hepatocellular carcinoma: current best practice and future direction. *Gastroenterology* **2019**;157(1):54-64.

10. Singal AG, Pillai A, Tiro J. Early detection, curative treatment, and survival rates for
15 hepatocellular carcinoma surveillance in patients with cirrhosis: a meta-analysis. *PLoS medicine* **2014**;11(4):e1001624.

11. Singal AG, Yopp A, Skinner CS, Packer M, Lee WM, Tiro JA. Utilization of hepatocellular carcinoma surveillance among American patients: a systematic review. *Journal of general internal medicine* **2012**;27(7):861-7.

20 12. Singal AG, Li X, Tiro J, Kandunoori P, Adams-Huet B, Nehra MS, *et al.* Racial, social, and clinical determinants of hepatocellular carcinoma surveillance. *The American journal of medicine* **2015**;128(1):90. e1-. e7.

13. Tzartzeva K, Obi J, Rich NE, Parikh ND, Marrero JA, Yopp A, *et al.* Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with
25 cirrhosis: a meta-analysis. *Gastroenterology* **2018**;154(6):1706-18. e1.

14. Benesova L, Belsanova B, Suchanek S, Kopeckova M, Minarikova P, Lipska L, *et al.* Mutation-based detection and monitoring of cell-free tumor DNA in peripheral blood of cancer patients. *Analytical biochemistry* **2013**;433(2):227-34.

15. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, *et al.* Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* **2015**;112(11):E1317-25 doi 10.1073/pnas.1500076112.
16. Cai J, Chen L, Zhang Z, Zhang X, Lu X, Liu W, *et al.* Genome-wide mapping of 5-hydroxymethylcytosines in circulating cell-free DNA as a non-invasive approach for early detection of hepatocellular carcinoma. *Gut* **2019**;68(12):2195-205 doi 10.1136/gutjnl-2019-318882.
17. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, *et al.* Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* **2017**;16(11):1155-61 doi 10.1038/nmat4997.
18. Wang Y, Zhou K, Wang X, Liu Y, Guo D, Bian Z, *et al.* Multiple-level copy number variations in cell-free DNA for prognostic prediction of HCC with radical treatments. *Cancer Sci* **2021**;112(11):4772-84 doi 10.1111/cas.15128.
19. Kisiel JB, Dukek BA, R VSRK, Ghaz HM, Yab TC, Berger CK, *et al.* Hepatocellular Carcinoma Detection by Plasma Methylated DNA: Discovery, Phase I Pilot, and Phase II Clinical Validation. *Hepatology* **2019**;69(3):1180-92 doi 10.1002/hep.30244.
20. Chalasani NP, Ramasubramanian TS, Bhattacharya A, Olson MC, Edwards VD, Roberts LR, *et al.* A Novel Blood-Based Panel of Methylated DNA and Protein Markers for Detection of Early-Stage Hepatocellular Carcinoma. *Clin Gastroenterol Hepatol* **2021**;19(12):2597-605 e4 doi 10.1016/j.cgh.2020.08.065.
21. Klein E, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology* **2021**;32(9):1167-77.
22. Costo de la Examinación del Cáncer (*Cancer Screening Cost*) con Galleri. <<https://www.galleri.com/the-galleri-test/cost>>.
23. Chalasani NP, Porter K, Bhattacharya A, Book AJ, Neis BM, Xiong KM, *et al.* Validation of a Novel Multitarget Blood Test Shows High Sensitivity to Detect Early Stage Hepatocellular Carcinoma. *Clin Gastroenterol H* **2022**;20(1):173-82. e7.
24. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat*

Commun **2021**;12(1):5060 doi 10.1038/s41467-021-24994-w.

25. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **2019**;570(7761):385-9 doi 10.1038/s41586-019-1272-6.

5 26. Zhang X, Wang Z, Tang W, Wang X, Liu R, Bao H, *et al.* Ultra-Sensitive and Affordable Assay for Early Detection of Primary Liver Cancer Using Plasma cfDNA Fragmentomics. *Hepatology* **2021**.

27. Chen L, Abou-Alfa GK, Zheng B, Liu JF, Bai J, Du LT, *et al.* Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in
10 cirrhotic patients. *Cell Res* **2021**;31(5):589-92 doi 10.1038/s41422-020-00457-7.

28. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, *et al.* Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *P Natl Acad Sci USA* **2018**;115(46):E10925-E33 doi 10.1073/pnas.1814616115.

15 29. Fortin J-P, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome biology* **2015**;16(1):1-23.

30. Choi JK, Kim YJ. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nat Genet* **2009**;41(4):498-503 doi 10.1038/ng.319.

31. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, *et al.* Inference of transcription
20 factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun* **2019**;10(1):4666 doi 10.1038/s41467-019-12714-4.

32. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an *In vivo* Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **2016**;164(1-2):57-68 doi 10.1016/j.cell.2015.11.050.

25 33. Cheneby J, Menetrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, *et al.* ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res* **2020**;48(D1):D180-D8 doi 10.1093/nar/gkz945.

34. Bejjani F, Evanno E, Zibara K, Piechaczyk M, Jariel-Encontre I. The AP-1

transcriptional complex: Local switch or remote command? *Biochim Biophys Acta Rev Cancer* **2019**;1872(1):11-23 doi 10.1016/j.bbcan.2019.04.003.

35. Gozdecka M, Lyons S, Kondo S, Taylor J, Li Y, Walczynski J, *et al.* JNK suppresses tumor formation via a gene-expression program mediated by ATF2. *Cell Rep* **2014**;9(4):1361-74 doi 10.1016/j.celrep.2014.10.043.

36. Yan P, Zhou B, Ma Y, Wang A, Hu X, Luo Y, *et al.* Tracking the important role of JUNB in hepatocellular carcinoma by single-cell sequencing analysis. *Oncol Lett* **2020**;19(2):1478-86 doi 10.3892/ol.2019.11235.

37. Coto-Llerena M, Tosti N, Taha-Mehlitz S, Kancharla V, Paradiso V, Gallon J, *et al.* Transcriptional Enhancer Factor Domain Family member 4 Exerts an Oncogenic Role in Hepatocellular Carcinoma by Hippo-Independent Regulation of Heat Shock Protein 70 Family Members. *Hepatol Commun* **2021**;5(4):661-74 doi 10.1002/hep4.1656.

38. Zhang Z, Fang X, Xie G, Zhu J. GATA3 is downregulated in HCC and accelerates HCC aggressiveness by transcriptionally inhibiting slug expression Corrigendum in/10.3892/ol.2021.12836. *Oncology letters* **2021**;21(3):1-.

39. Zhang X, Hua L, Yan D, Zhao F, Liu J, Zhou H, *et al.* Overexpression of PCBP2 contributes to poor prognosis and enhanced cell growth in human hepatocellular carcinoma. *Oncol Rep* **2016**;36(6):3456-64.

40. Xiang X, Fu Y, Zhao K, Miao R, Zhang X, Ma X, *et al.* Cellular senescence in hepatocellular carcinoma induced by a long non-coding RNA-encoded peptide PINT87aa by blocking FOXM1-mediated PHB2. *Theranostics* **2021**;11(10):4929-44 doi 10.7150/thno.55672.

41. Shen M, Li S, Zhao Y, Liu Y, Liu Z, Huan L, *et al.* Hepatic ARID3A facilitates liver cancer malignancy by cooperating with CEP131 to regulate an embryonic stem cell-like gene signature. *Cell Death & Disease* **2022**;13(8):1-13.

42. Marchio A, Pineau P, Meddeb M, Terris B, Tiollais P, Bernheim A, *et al.* Distinct chromosomal abnormality pattern in primary liver cancer of non-B, non-C patients. *Oncogene* **2000**;19(33):3733-8 doi 10.1038/sj.onc.1203713.

43. Longerich T, Mueller MM, Breuhahn K, Schirmacher P, Benner A, Heiss C. Oncogenetic tree modeling of human hepatocarcinogenesis. *Int J Cancer* **2012**;130(3):575-83.

44. Stewart SL, Kwong SL, Bowlus CL, Nguyen TT, Maxwell AE, Bastani R, *et al.* Racial/ethnic disparities in hepatocellular carcinoma treatment and survival in California, 1988-2012. *World journal of gastroenterology* **2016**;22(38):8584-95 doi 10.3748/wjg.v22.i38.8584.
45. Gupta S, Bent S, Kohlwes J. Test characteristics of alpha-fetoprotein for detecting
5 hepatocellular carcinoma in patients with hepatitis C. A systematic review and critical analysis. *Ann Intern Med* **2003**;139(1):46-50 doi 10.7326/0003-4819-139-1-200307010-00012.
46. Zhao C, Jin M, Le RH, Le MH, Chen VL, Jin M, *et al.* Poor adherence to
hepatocellular carcinoma surveillance: a systematic review and meta-analysis of a complex
issue. *Liver International* **2018**;38(3):503-14.
47. Bokhorst LP, Alberts AR, Rannikko A, Valdaghi R, Pickles T, Kakehi Y, *et al.*
10 Compliance Rates with the Prostate Cancer Research International Active Surveillance
(PRIAS) Protocol and Disease Reclassification in Noncompliers. *Eur Urol* **2015**;68(5):814-21
doi 10.1016/j.eururo.2015.06.012.
48. Duffy MJ, van Rossum LG, van Turenhout ST, Malminiemi O, Sturgeon C, Lamerz
15 R, *et al.* Use of faecal markers in screening for colorectal neoplasia: a European group on
tumor markers position paper. *Int J Cancer* **2011**;128(1):3-11.
49. Singal AG, Lampertico P, Nahon P. Epidemiology and surveillance for
hepatocellular carcinoma: New trends. *J Hepatol* **2020**;72(2):250-61.
50. Heimbach JK, Kulik LM, Finn RS, Sirlin CB, Abecassis MM, Roberts LR, *et al.*
20 AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology* **2018**;67(1):358-
80 doi 10.1002/hep.29086.
51. Zhang B-H, Yang B-H, Tang Z-Y. Randomized controlled trial of screening for
hepatocellular carcinoma. *Journal of cancer research and clinical oncology* **2004**;130(7):417-
22.
52. Goh SK, Do H, Testro A, Pavlovic J, Vago A, Lokan J, *et al.* The Measurement of
25 Donor-Specific Cell-Free DNA Identifies Recipients With Biopsy-Proven Acute Rejection
Requiring Treatment After Liver Transplantation. *Transplant Direct* **2019**;5(7):e462 doi
10.1097/txd.0000000000000902.
53. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
30 *Bioinformatics* **2018**;34(17):i884-i90 doi 10.1093/bioinformatics/bty560.

54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012**;9(4):357-9 doi 10.1038/nmeth.1923.

55. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **2015**;31(12):2032-4 doi 10.1093/bioinformatics/btv098.

56. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**;26(6):841-2 doi 10.1093/bioinformatics/btq033.

57. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **2017**;8(1):1324 doi 10.1038/s41467-017-00965-y.

58. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *Bmc Bioinformatics* **2011**;12:77 doi 10.1186/1471-2105-12-77.

59. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **2017**;355(6322):eaaf8399 doi 10.1126/science.aaf8399.

60. Bendix Carstensen M, Esa L, Michael H. Epi: A package for statistical analysis in epidemiology. Versión del paquete R **2016**;20.

Otras modalidades

Aunque la divulgación se ha descrito junto con la descripción detallada de la misma, la descripción anterior pretende ilustrar y no limitar el alcance de la divulgación, que se define por el alcance de las reivindicaciones adjuntas. Otros aspectos, ventajas, y modificaciones están dentro del alcance de las siguientes reivindicaciones.

La literatura científica y de patentes a la que se hace referencia en la presente establece los conocimientos de que disponen los expertos en la materia. Todas las patentes de EE.UU. y solicitudes de patente de EE.UU. publicadas o no publicadas citadas en la presente se incorporan por referencia. Todas las patentes extranjeras publicadas y las solicitudes de patente citadas en la presente se incorporan por referencia. Todas las demás referencias publicadas, documentos, manuscritos y literatura científica citados en la presente se incorporan por referencia.