(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 882 293**

(21) Número de solicitud: 202030506

(51) Int. CI.:

C12Q 1/6886 (2008.01)
G01N 33/574 (2006.01)

(12) PATENTE DE INVENCIÓN CON EXAMEN B2

(22) Fecha de presentación:

01.06.2020

(43) Fecha de publicación de la solicitud:

01.12.2021

Fecha de modificación de las reivindicaciones:

01.03.2022

Fecha de concesión:

17.11.2023

(45) Fecha de publicación de la concesión:

24.11.2023

(73) Titular/es:

SERVICIO ANDALUZ DE SALUD (75.0%)
Avda. de la Constitución, 18
41071 Sevilla (Sevilla) ES y
UNIVERSIDAD DE MÁLAGA (25.0%)

(72) Inventor/es:

GÁLVEZ CARVAJAL, Laura;
SÁNCHEZ MUÑOZ, Alfonso y
ALBA CONEJO, Emilio

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

Observaciones:

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

(54) Título: **MÉTODO PARA PREDECIR O PRONOSTICAR EL RIESGO DE RECAÍDA EN PACIENTES CON TUMORES DE CÉLULAS GERMINALES NO SEMINOMA TRAS ORQUIECTOMÍA**

(57) Resumen:

La presente invención se refiere a un método in vitro de obtención de datos útiles para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía. El método comprende medir en una muestra biológica aislada del individuo el producto de expresión de MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16. También se describe un kit de diagnóstico para llevar a cabo este método y sus usos médicos.

Fig. 1

DESCRIPCIÓN

MÉTODO PARA PREDECIR O PRONOSTICAR EL RIESGO DE RECAÍDA EN PACIENTES CON TUMORES DE CÉLULAS GERMINALES NO SEMINOMA TRAS ORQUIECTOMÍA

CAMPO DE LA INVENCIÓN

La presente invención se encuentra dentro del campo de la medicina de precisión, y se refiere a un método de obtención de datos útiles para predecir o pronosticar el riesgo de recaída en pacientes con tumores de células germinales no seminoma tras orquiectomía.

ESTADO DE LA TÉCNICA ANTERIOR

Actualmente para el 15-20% de los pacientes con tumores de células germinales (TCG) mestastásicos las opciones curativas son limitadas, estando en investigación la respuesta a terapias diana (1). Como estrategias postorquiectomía en los pacientes con tumores de células germinales no seminoma (TCGNS) estadio I son válidas la quimioterapia adyuvante, la linfadenectomía retroperitoneal y la vigilancia activa. La infiltración linfovascular es el factor pronóstico más aceptado y asociado al riesgo de recaída en TCGNS (aumentándolo hasta en un 50%) (2–5). El tratamiento quimioterápico adyuvante en estos pacientes con estadio I basado en la presencia de infiltración linfovascular supone el sobretratamiento del restante 50%, con la consiguiente toxicidad aguda y a largo plazo, por lo que se requieren marcadores más fidedignos de recaída. En los tumores de células germinales testiculares (TCGT) los estudios de expresión génica se encuentran en etapas muy iniciales de investigación.

DESCRIPCIÓN DE LAS FIGURAS

Fig. 1. Representación de los pasos del procesamiento de las muestras: obtención del bloque tumoral, tinción hematoxilina-eosina para identificar el área tumoral, marcaje del área tumoral en áreas no teñidas para la macrodisección y extracción del ARN.

Fig. 2. **Visión general del análisis de expresión génica con nCounter™**:
1) Esquema de la sonda portadora del objetivo específico (target-specific reporter probe) y la sonda de captura (target-specific capture probe). Proceso de hibridación a 65ºC durante 12-21 horas en placa térmica.

2) Estación de preparación de muestras donde se eliminan los excesos de sondas y se alinean en el cartucho los híbridos (sonda + ARN diana).

3) Estación de análisis digital, donde se procede al contaje digital de moléculas.

Fig. 3. Gráfico tipo volcán. Representa la expresión diferencial de los genes tras el análisis de las 54 muestras tumorales de pacientes con TCGNS, incluyendo tanto a pacientes con recaída como sin recaída. El eje Y representa la pvalue ajustado al total de genes estudiados, considerándose < 0.05 significativo. El eje X representa el log_2 fold change, expresando el número de veces que cada gen es sobreexpresado o infraexpresado en el grupo con recaída respecto a los pacientes sin recaída. Se han delimitado los límites -1 y 1 (doble infraexpresado o doble sobreexpresado). No se observa ningún gen con expresión diferencial estadísticamente significativa y la gran mayoría de los genes tenían una expresión tan sólo entre 1 y -1 fold change.

Fig. 4. Representación del número de genes seleccionados con cada uno de los criterios en función del tamaño umbral empleado. Azul: genes seleccionados según el criterio de voto mayoritario, en función del tamaño umbral empleado. Rojo: genes seleccionados según el criterio de unanimidad, en función del tamaño umbral empleado.

DESCRIPCIÓN DE LA INVENCIÓN

Los autores de la presente invención han desarrollado un método que permite predecir o pronosticar el riesgo de recaída en pacientes con tumores de células germinales no seminoma tras orquiectomía. Han estudiado 730 genes en cada uno de los 54 pacientes incluidos en el estudio, identificando aquellos que son relevantes para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma tras orquiectomía.

USO DE LOS MARCADORES DE LA INVENCIÓN

Un **primer aspecto** de la invención se refiere a *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía.

Una realización preferida de este aspecto se refiere al uso por separado de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*, o cualquiera de sus combinaciones, para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma tras orquiectomía.

Otra realización preferida de refiere al uso simultáneo de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16,* para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma tras orquiectomía.

En esta memoria se entiende por *MPL proto-oncogene, thrombopoietin receptor* (también llamado MPLV; TPOR; C-MPL; CD110; THPOR; THCYT2) tanto al gen como a la proteína. En 1990, se identificó un oncogén, v-mpl, del virus de la leucemia mieloproliferativa murina que era capaz de inmortalizar las células hematopoyéticas de la médula ósea de diferentes linajes. En 1992 se clonó el homólogo humano, llamado c-mpl. Los datos de secuencia revelaron que c-mpl codificaba una proteína que era homóloga con miembros de la superfamilia de receptores hematopoyéticos. La presencia de oligodesoxinucleótidos antisentido de c-mpl inhibió la formación de colonias de megacariocitos. El ligando para c-mpl, la trombopoyetina, se clonó en 1994. Se demostró que la trombopoyetina es el principal regulador de la megacariocitopoyesis y la formación de plaquetas. La proteína codificada por el gen c-mpl, CD110, es un dominio transmembrana de 635 aminoácidos, con dos dominios extracelulares de receptor de citocina y dos motivos de caja de receptor de citocina intracelular. Los ratones con deficiencia de TPO-R fueron severamente trombocitopénicos, enfatizando el importante papel de CD110 y trombopoyetina en la formación de megacariocitos y plaquetas. Tras la unión de la trombopoyetina, el CD110 se dimeriza y la familia JAK de tirosina quinasas no receptoras, así como la familia STAT, la familia MAPK, la proteína adaptadora Shc y los receptores mismos se fosforilan.

En el contexto de la presente invención, *MPL* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *MLP*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 1,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 1. en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *MPL*. Preferiblemente, es la SEQ ID NO: 2

En esta memoria se entiende por *LAMB4 laminin subunit beta 4* tanto el gen como la proteína.

En el contexto de la presente invención, *LAMB4* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *LAMB4*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 3,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 3. en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *LAMB4*. Preferiblemente, es la SEQ ID NO: 4

En esta memoria se entiende por *FANCB* o *FA complementation group B* (también llamado FA2; FAB; FACB; FAAP90; FAAP95) tanto al gen como a la proteína. Este gen codifica a un miembro del grupo B de complementación de anemia de Fanconi. Esta proteína se ensambla en un complejo de nucleoproteína que está involucrado en la reparación de las lesiones de ADN. Las mutaciones en este gen pueden causar inestabilidad cromosómica y síndrome de VACTERL con hidrocefalia.

En el contexto de la presente invención, *FANCB* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *FANCB*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 5,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 5. en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *FANCB*. Preferiblemente, es la SEQ ID NO: 6

En esta memoria se entiende por *NR4A1* o *nuclear receptor subfamily 4 group A member 1* (también llamado HMR; N10; TR3; NP10; GFRP1; NAK-1; NGFIB; NUR77) tanto al gen como a la proteína. Este gen codifica a un miembro de la superfamilia de receptores de esteroides-hormona tiroidea-retinoide. La expresión es inducida por fitohemaglutinina en linfocitos humanos y por la estimulación sérica de fibroblastos detenidos. La proteína codificada actúa como un factor de transcripción nuclear. La translocación de la proteína del núcleo a las mitocondrias induce apoptosis. Se han encontrado múltiples variantes de transcripción que codifican diferentes isoformas para este gen.

En el contexto de la presente invención, *NR4A1* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *NR4A1*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 7,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 7. en las que el polipéptido codificado por dichos ácidos

nucleicos posee la actividad y las características estructurales de la proteína *NR4A1*. Preferiblemente, es la SEQ ID NO: 8

En esta memoria se entiende por *CREB5* o *cAMP responsive element binding protein 5* (también llamado CREB-5; CREBPA; CRE-BPA) tanto al gen como a la proteína. El producto de este gen pertenece a la familia de proteínas de unión a CRE (elemento de respuesta de AMPc). Los miembros de esta familia contienen dominios de unión a ADN de zinc-dedo y bZIP. La proteína codificada se une específicamente a CRE como un homodímero o un heterodímero con c-Jun o CRE-BP1, y funciona como un activador trans dependiente de CRE. Alternativamente, se han identificado variantes de transcripción empalmadas que codifican diferentes isoformas.

En el contexto de la presente invención, *CREB5* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *CREB5*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 9,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 9. en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *CREB5*. Preferiblemente, es la SEQ ID NO: 10.

En esta memoria se entiende por *CALML3 o calmodulin like 3* (también llamado CLP) tanto al gen como a la proteína.

En el contexto de la presente invención, *CALML3* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *CALML3*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 11,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 11. en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *CREB5*. Preferiblemente, es la SEQ ID NO: 12.

En esta memoria se entiende por *FOSL1, FOS like 1 o AP-1 transcription factor subunit* (también llamado FRA; FRA1; fra-1) tanto al gen como a la proteína. La familia de genes Fos consta de 4 miembros: FOS, FOSB, FOSL1 y FOSL2. Estos genes codifican proteínas de cremallera de leucina que pueden dimerizarse con proteínas de la familia JUN, formando así el complejo del factor de transcripción AP-1. Como tal, las proteínas FOS se han implicado como reguladores de la proliferación, diferenciación y transformación celular. Se han encontrado varias variantes de transcripción que codifican diferentes isoformas para este gen.

En el contexto de la presente invención, *FOSL1* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *FOSL1*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 13,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 13, en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *CREB5*. Preferiblemente, es la SEQ ID NO: 14.

En esta memoria se entiende por *COL24A1* o *collagen type XXIV alpha 1 chain* tanto al gen como a la proteína. Este gen es miembro de la familia de genes de colágeno y se cree que regula la fibrilogénesis de colágeno tipo I durante el desarrollo fetal.

En el contexto de la presente invención, *COL24A1* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *COL24A1*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 15,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 15, en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *CREB5*. Preferiblemente, es la SEQ ID NO: 16.

En esta memoria se entiende por *ZBTB16* o *zinc finger and BTB domain containing 16* (también llamado PLZF; ZNF145) tanto al gen como a la proteína. Este gen es miembro de la familia de proteínas de dedos de zinc tipo Krueppel C2H2 y codifica un factor de transcripción de dedos de zinc que contiene nueve dominios de dedos de zinc tipo Kruppel en el extremo carboxilo. Esta proteína se encuentra en el núcleo, participa en la progresión del ciclo celular e interactúa con una histona desacetilasa. Los casos específicos de reordenamiento genético aberrante en este locus se han asociado con leucemia promielocítica aguda (APL). Se han caracterizado variantes de empalme transcripcional alternativas.

En el contexto de la presente invención, *ZBTB16* se define también por una secuencia de nucleótidos o polinucleótido, que constituye la secuencia codificante de la proteína *ZBTB16*, y que comprendería diversas variantes procedentes de:

a) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica de la SEQ ID NO: 17,

b) moléculas de ácido nucleico cuya cadena complementaria híbrida con la secuencia polinucleotídica de a),

c) moléculas de ácido nucleico cuya secuencia difiere de a) y/o b) debido a la degeneración del código genético,

d) moléculas de ácido nucleico que codifican un polipéptido que comprende la secuencia aminoacídica con una identidad de al menos un 80%, un 90%, un 95%, un 98% o un 99% con la SEQ ID NO: 17, en las que el polipéptido codificado por dichos ácidos nucleicos posee la actividad y las características estructurales de la proteína *ZBTB16*. Preferiblemente, es la SEQ ID NO: 18.

MÉTODO PARA PREDECIR O PRONOSTICAR EL RIESGO DE RECAÍDA

Otro **aspecto** de la invención se refiere a un método *in vitro* de obtención de datos útiles para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma tras orquiectomía, de ahora en adelante primer método de la invención, que comprende medir en una muestra biológica aislada del individuo el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*.

En una realización preferida de este aspecto de la invención, el primer método de la invención además comprende comparar las cantidades obtenidas con una cantidad de referencia.

Otro a**specto** de la invención se refiere a un método *in vitro* para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía, de ahora en adelante segundo método de la invención, que comprende medir en una muestra biológica aislada el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16,* comparar las cantidades obtenidas con una cantidad de referencia, e incluir al individuo en el grupo de individuos con una mayor probabilidad de recaída, cuando

a) MLP está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

b) *LAMB4* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

c) *FANCB* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

d) *NR4A1* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

e) *CREB5* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

f) *CALML3* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

g) *FOSL1* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

h) *COL24A1* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

i) *ZBTB16* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

Los pasos de medida del producto de expresión de los marcadores y de su comparación con una muestra de referencia del método descrito anteriormente pueden ser total o parcialmente automatizados, por ejemplo, por medio de un equipo robótico sensor para la detección del producto de expresión de los marcadores o la clasificación computarizada en los distintos grupos.

El término "comparación", tal y como se utiliza en la descripción, se refiere pero no se limita, a la comparación del resultado de la muestra biológica a analizar, también llamada muestra biológica problema, con una amplificación de una o varias muestras de referencia deseable. La muestra de referencia puede ser analizada, por ejemplo, simultánea o consecutivamente, junto con la muestra biológica problema. Las comparaciones descritas en los métodos de la presente invención pueden ser realizadas manualmente o asistida por ordenador.

En la presente invención "pronóstico" se entiende como la evolución esperada de una enfermedad y se refiere a la valoración de la probabilidad según la cual un sujeto padece una enfermedad así como a la valoración de su inicio, estado de desarrollo, evolución, o de su regresión, y/o el pronóstico del curso de la enfermedad en el futuro. Como entenderán los expertos en la materia, tal valoración, aunque se prefiere que sea, normalmente puede no ser correcta para el 100% de los sujetos que se va a diagnosticar. El término, sin embargo, requiere que una parte estadísticamente significativa de los sujetos se pueda identificar como que padecen la enfermedad o que tienen predisposición a la misma. Si una parte es estadísticamente significativa se puede determinar sin más por el experto en la materia usando varias herramientas de evaluación estadística bien conocidas, por ejemplo, determinación de intervalos de confianza, determinación de valores p, prueba t de Student, prueba de Mann-Whitney, etc. Los intervalos de confianza preferidos son al menos el 50%, al menos el 60%, al menos el 70%, al menos el 80%, al menos el 90%, al menos el 95%. Los valores de p son, preferiblemente, 0,2, 0,1, 0,05.

Por "predicción de la respuesta" se entiende, en el contexto de la presente invención, la determinación de la probabilidad de que el paciente responda de forma favorable o desfavorable a una terapia o a un tratamiento determinado. Especialmente, el término "predicción", como se usa aquí, se refiere a una evaluación individual de cualquier parámetro que pueda ser útil en determinar la evolución de un paciente. Como entenderán los expertos en la materia, la predicción de la respuesta clínica al tratamiento, aunque se prefiere que sea, no necesita ser correcta para el 100% de los sujetos a ser diagnosticados o evaluados. El término, sin embargo, requiere que se pueda identificar una parte estadísticamente significativa de los sujetos como que tienen una probabilidad aumentada de tener una respuesta positiva. El experto en la materia puede determinar fácilmente si un sujeto es estadísticamente significativo usando varias herramientas de evaluación estadística bien conocidas, por ejemplo, determinación de intervalos de confianza, determinación de los valores de p, prueba t de Student, prueba de Mann Whitney, etc. Los intervalos de confianza preferidos son al menos del 50%>, al menos del 60%>, al menos del 70%>, al menos del 80%>, al menos del 90%), al menos del 95%>. Preferiblemente, el valor de p es menor de 0,1 , de 0,05, de 0,01 , de 0,005 o de 0,0001. Preferiblemente, la presente invención permite clasificar correctamente a los individuos de forma diferencial en al menos el 60%, más preferiblemente en al menos el 70%, mucho más preferiblemente en al menos el 80%, o aún mucho más preferiblemente en al menos el 90% de los sujetos de un determinado

grupo o población analizada. La predicción de la respuesta clínica se puede hacer utilizando cualquier criterio de valoración usado y conocido por el experto en la materia.

Una "muestra biológica aislada" incluye, pero sin limitarnos a, células, tejidos y/o fluidos biológicos de un organismo, obtenidos mediante cualquier método conocido por un experto en la materia. Preferiblemente, la muestra biológica aislada es tejido.

En una realización preferida de este aspecto de la invención la muestra biológica es una muestra de tejido, preferiblemente, una muestra de tejido del tumor primario.

El término "individuo", tal y como se utiliza en la descripción, se refiere a animales, preferiblemente mamíferos, y más preferiblemente, humanos. El término "individuo" en esta memoria, es sinónimo de "paciente", y no pretende ser limitativo en ningún aspecto, pudiendo ser éste de cualquier edad, sexo y condición física.

En la invención, el método para determinar el resultado, es decir, el nivel de expresión, no necesita estar particularmente limitado, y puede seleccionarse mediante un método de perfilado de genes, como una micromatriz, y/o un método que comprende PCR, tal como tiempo de PCR; y/o Northern Blot. La PCR cuantitativa en tiempo real (generalmente abreviada como RQ-PCR, RT-qPCR, rt-PCR o qPCR) es una técnica de cuantificación de la expresión génica sensible y reproducible que se puede usar particularmente para perfilar la expresión de miARN en células y tejidos. Se puede utilizar cualquier método para evaluar los resultados de la RT-PCR, y se puede preferir el método ΔΔCt. El método ΔΔCt se describe en detalle por Livak et al. (Methods 2001, 25: 402-408). (Ct = Valores umbral de ciclo).

En otra realización preferida la determinación del producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* se realiza mediante:

a) un procedimiento de análisis de la expresión génica, como, por ejemplo, pero sin limitarnos, qRT-PCR, microarrays de ADN, nCounter, RNA-Seq, FISH, microarrays de tejidos o transferencia Northern; y/o

b) un inmunoensayo o inmunohistoquímica.

El ensayo nCounter se basa en la detección digital directa de las moléculas de ARNm de interés utilizando pares de sondas codificadas por colores específicos del objetivo. No requiere la conversión de ARNm a ADNc por transcripción inversa o la amplificación del ADNc resultante por PCR. Cada gen objetivo de interés se detecta

usando un par de sondas de captura y un reportero que llevan secuencias específicas de la diana de 35 a 50 bases. Adicionalmente, cada sonda indicadora lleva un código de color único en el extremo 5 'que permite el código de barras molecular de los genes de interés, mientras que las sondas de captura llevan una etiqueta de biotina en el extremo 3' que proporciona un asa molecular para la unión de genes diana para facilitar detección digital aguas abajo. Después de la hibridación en fase de solución entre el ARNm objetivo y los pares de sondas de captura de reportero, se eliminan las sondas en exceso y los complejos sonda / objetivo se alinean e inmovilizan en el cartucho nCounter, que luego se coloca en un analizador digital para la adquisición de imágenes y el procesamiento de datos. Cientos de miles de códigos de color que designan objetivos de ARNm de interés se graban directamente en la superficie del cartucho. El nivel de expresión de un gen se mide contando el número de veces que se detecta el código de barras con código de color para ese gen.

En otra realización aun más preferida la determinación del producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* se realiza mediante nCounter.

USOS MÉDICOS DE LA INVENCIÓN

La invención permite la posibilidad de seleccionar qué pacientes con tumores de células germinales no seminoma requieren un tratamiento adyuvante tras la orquiectomía, por presentar según este modelo de nueve genes un alto riesgo de recaída. Este tratamiento consistiría en quimioterapia adyuvante tipo BEP (bleomicina, etopósido, cisplatino) en número de uno o dos ciclos.

Por tanto, otro aspecto de la invención se refiere a quimioterapia adyuvante tipo BEP (bleomicina, etopósido, cisplatino) para el tratamiento de un individuo clasificado en el grupo de individuos con una mayor probabilidad de recaída según cualquiera de los métodos de la invención.

Permite además seleccionar aquellos pacientes con tumores de células germinales no seminoma de bajo riesgo de recaída tras orquiectomía, los cuales no precisarían ningún tipo de tratamiento añadido a la cirugía del tumor primario. Eliminando, por tanto, la toxicidad aguda y a largo plazo del esquema BEP en estos pacientes, que no haría más que aportarle toxicidad.

Además supondría un gasto eficiente de los fondos públicos monetarios, con una distribución más acertada del gasto.

KIT O DISPOSITIVO DE LA INVENCIÓN

Otro aspecto de la invención se refiere a un kit o dispositivo, de ahora en adelante kit o dispositivo de la invención, que comprende los elementos y/o reactivos necesarios para cuantificar el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*.

En una realización preferida, el kit o dispositivo de la invención comprende cebadores, sondas y/o anticuerpos capaces de cuantificar el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*, y donde:

- los cebadores o primers son secuencias de polinucleótidos de entre 10 y 30 pares de bases, más preferiblemente de entre 15 y 25 pares de bases, aún más preferiblemente de entre 18 y 22 pares de bases, y aún mucho más preferiblemente de alrededor de 20 pares de bases, que presentan una identidad de al menos un 80%, más preferiblemente de al menos un 90%, aún más preferiblemente de al menos un 95%, aún mucho más preferiblemente de al menos un 98%, y particularmente de un 100%, con un fragmento de la secuencia complementaria a la SEQ ID NO: 2, SEQ ID NO 4, SEQ ID NO: 6, SEQ ID NO: 8, SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14, SEQ ID NO: 16, SEQ ID NO: 18,

las sondas son secuencias de polinucleótidos de entre 30 y 1100 pares de bases, más preferiblemente de entre 100 y 1000 pares de bases, y aún más preferiblemente de entre 150 y 500 pares de bases, que presentan una identidad de al menos un 80%, más 25 preferiblemente de al menos un 90%, aún más preferiblemente de al menos un 95%, aún mucho más preferiblemente de al menos un 98%, y particularmente de un 100%, con un fragmento de las secuencias complementarias a la SEQ ID NO: 2, SEQ ID NO 4, SEQ ID NO: 6, SEQ ID NO: 8, SEQ ID NO: 10, SEQ ID NO: 12, SEQ ID NO: 14, SEQ ID NO: 16, SEQ ID NO: 18,

- los anticuerpos son capaces de unirse específicamente a una región formada por la secuencia aminoacídica SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5, SEQ ID NO: 7, SEQ ID NO: 9, SEQ ID NO: 11, SEQ ID NO: 13, SEQ ID NO: 15, SEQ ID NO: 17.

El kit de la invención puede incluir controles positivos y/o negativos. El kit además puede contener, sin ningún tipo de limitación, tampones, soluciones de extracción de ADN/ARN o proteínas, agentes para prevenir la contaminación, inhibidores de la degradación del ADN/ARN y/o las proteínas, etc.

Por otro lado el kit puede incluir todos los soportes y recipientes necesarios para su puesta en marcha y optimización. Preferiblemente, el kit comprende además las instrucciones para llevar a cabo los métodos de la invención.

El kit o dispositivo de la invención puede usarse y el uso no está particularmente limitado, aunque se prefiere el uso en el método de la invención en cualquiera de sus realizaciones.

Por tanto, otro **aspecto** de la invención se refiere al uso de un kit o dispositivo de la invención para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma tras orquiectomía.

En otra realización preferida de este aspecto de la invención, los oligonucleótidos, cebadores, sondas o anticuerpos están modificados o marcados, por ejemplo, pero sin limitarnos, mediante un marcaje radiactivo o inmunológico. Así, preferiblemente, los oligonucleótidos presentan modificaciones en alguno de sus nucleótidos, como por ejemplo, pero sin limitarnos a, nucleótidos que tengan alguno de sus átomos con un isótopo radiactivo, normalmente ^{32}P o tritio, nucleótidos marcados inmunológicamente, como por ejemplo con una molécula de digoxigenina, y/o inmovilizadas en una membrana. Varias posibilidades son conocidas en el estado de la técnica.

AUTOMATIZACIÓN DE LOS MÉTODOS DE LA INVENCIÓN

La invención se extiende también a programas de ordenador adaptados para que cualquier medio de procesamiento pueda llevar a la práctica los métodos de la invención. Tales programas pueden tener la forma de código fuente, código objeto, una fuente intermedia de código y código objeto, por ejemplo, como en forma parcialmente compilada, o en cualquier otra forma adecuada para uso en la puesta en práctica de los procesos según la invención. Los programas de ordenador también abarcan aplicaciones en la nube basadas en dicho procedimiento.

En particular, la invención abarca programas de ordenador dispuestos sobre o dentro de una portadora. La portadora puede ser cualquier entidad o dispositivo capaz de soportar el programa. Cuando el programa va incorporado en una señal que puede ser transportada directamente por un cable u otro dispositivo o medio, la portadora puede estar constituida por dicho cable u otro dispositivo o medio. Como variante, la portadora podría ser un circuito integrado en el que va incluido el programa, estando el circuito integrado adaptado para ejecutar, o para ser utilizado en la ejecución de, los procesos correspondientes.

Por ejemplo, los programas podrían estar incorporados en un medio de almacenamiento, como una memoria ROM, una memoria CD ROM o una memoria ROM de semiconductor, una memoria USB, o un soporte de grabación magnética, por ejemplo, un disco flexible o un disco duro. Alternativamente, los programas podrían estar

soportados en una señal portadora transmisible. Por ejemplo, podría tratarse de una señal eléctrica u óptica que podría transportarse a través de cable eléctrico u óptico, por radio o por cualesquiera otros medios.

La invención se extiende también a programas de ordenador adaptados para que cualquier medio de procesamiento pueda llevar a la práctica los métodos de la invención. Tales programas pueden tener la forma de código fuente, código objeto, una fuente intermedia de código y código objeto, por ejemplo, como en forma parcialmente compilada, o en cualquier otra forma adecuada para uso en la puesta en práctica de los procesos según la invención. Los programas de ordenador también abarcan aplicaciones en la nube basadas en dicho procedimiento.

Por tanto, otro aspecto de la invención se refiere a un medio de almacenamiento legible por un ordenador que comprende instrucciones de programa capaces de hacer que un ordenador lleve a cabo los pasos de cualquiera de los métodos de la invención.

Otro aspecto de la invención se refiere a una señal transmisible que comprende instrucciones de programa capaces de hacer que un ordenador lleve a cabo los pasos de cualquiera de los métodos de la invención.

Los términos "polinucleótido" y "ácido nucleico" se usan aquí de manera intercambiable, refiriéndose a formas poliméricas de nucleótidos de cualquier longitud, tanto ribonucleótidos (ARN ó RNA) como desoxirribonucleótidos (ADN ó DNA). Los términos "secuencia aminoacídica", "péptido", "oligopéptido", "polipéptido" y "proteína" se usan aquí de manera intercambiable, y se refieren a una forma polimérica de aminoácidos de cualquier longitud, que pueden ser codificantes o no codificantes, química o bioquímicamente modificados.

A lo largo de la descripción y las reivindicaciones la palabra "comprende" y sus variantes no pretenden excluir otras características técnicas, aditivos, componentes o pasos. Para los expertos en la materia, otros objetos, ventajas y características de la invención se desprenderán en parte de la descripción y en parte de la práctica de la invención. Los siguientes ejemplos y dibujos se proporcionan a modo de ilustración, y no se pretende que sean limitativos de la presente invención.

EJEMPLOS DE LA INVENCIÓN

Materiales y métodos

Obtención y procesamiento de las muestras

Durante el primer año de trabajo fueron identificados todos los pacientes incluidos en el estudio y seleccionados los bloques de tumor de TCGNS estadio I representativos de la lesión, fijados en formol e incluidos en parafina (FFPE). Todos los bloques tumorales procedían de la pieza quirúrgica de la orquiectomía. Los casos de pacientes intervenidos en la provincia de Málaga, fueron solicitados y cedidos por la Unidad de Biobanco de IBIMA (Instituto de Investigación Biomédica de Málaga). En el caso de los distintos hospitales que participaron en este proyecto, el material solicitado fue el estrictamente necesario, de modo que no hubo ningún excedente y no fue necesario ningún plan de contingencia.

Una vez identificadas las muestras, en cada uno de los hospitales colaboradores, se realizó una tinción de hematoxilina-eosina a partir de la cual el patólogo delimitó la zona correspondiente al tumor. Esta zona se microdiseccionó manualmente, obteniendo 6 cortes de 10µm de grosor. Estos cortes se enviaron por mensajería al Laboratorio de Biología Molecular del Cáncer ubicado en el Centro de Investigaciones Médico Sanitarias (CIMES) de la Universidad de Málaga (UMA)), donde se llevó a cabo la extracción, purificación y cuantificación del ARN total de las muestras (*figura 1*). La extracción del ARN de las muestras se llevó a cabo a partir de estos cortes con el kit RNeasy FFPE (Qiagen). Posteriormente se determinó la concentración y calidad del ARN purificado con el espectrofotómetro Nanodrop (Thermo Scientific). El ARN se cuantificó y se almacenó a -80ºC en alícuotas de un solo uso.

Previo al procesamiento del ARN, se seleccionó las muestras de ARN que pasaban el criterio de calidad de Nanostring (concentración ≥ 12.5 ng/µL y ratio A260/280 de 1.7-2.3, medido con Nanodrop). El ARN se procesó siguiendo las recomendaciones y pautas del fabricante.

Estudio de expresión génica

Para el estudio de expresión génica se ha utilizado el sistema nCounter™ de Nanostring, basado en el contaje digital de moléculas de ARN, de manera automática y con capacidad para estudiar en una misma reacción hasta 800 dianas de interés.

Es una metodología fundamentada en la hibridación, sin amplificación previa, donde cada molécula diana se va a identificar por un código o combinación de seis fluorocromos de cuatro colores diferentes. La secuencia génica problema hibridará con dos sondas: una sonda de captura que está unida a biotina y una sonda reporter unida a la combinación de los seis fluorocromos.

Concretamente en este trabajo se ha empleado el panel nCounter™ PanCancer Pathways (Nanostring). Este panel incluye 730 genes esenciales directamente relacionados con las 13 rutas canónicas principales que gobiernan el proceso de transformación en cáncer, incluyendo genes claves relacionados con los procesos de proliferación celular, apoptosis, inestabilidad genómica y transición epitelio-mesénquima.

El proceso ha consistido en la hibridación de ARN, tras lo cual se prepararon las muestras para poder realizar el contaje digital de los genes objeto de estudio.

Las muestras se hibridaron con las sondas reporter y de captura incubándose a 65ºC toda la noche en una placa térmica, concretamente un termociclador, tiempo tras el cual las muestras se transfirieron a la estación de preparación o PrepStation. En esta, se eliminaron los excesos de sonda sin hibridar, se purificaron las muestras y se produjo la unión de las sondas hibridadas al cartucho. Las moléculas inmovilizadas fueran sometidas a un campo eléctrico para su posterior alineamiento.

Finalmente, en el módulo Digital Analyzer o estación de análisis digital, se colocaron los cartuchos para realizar el contaje directo de los códigos de barras asociados a los ARN mensajeros (*figura 2*).

El sistema dispone de un software libre para el análisis de los datos resultantes, el nSolver Analysis Software (versión 5). El panel PanCancer Pathways incorpora 40 genes de normalización ("*housekeeping genes*"), que fueron seleccionados de los estudios de expresión génica del TCGA (*The Cancer Genome Atlas*), tras demostrar su representatividad en diferentes tipos de cáncer. El método de ajuste de p-value fue por el método de Bonferroni. Se usó este software para la normalización de los datos y para realizar los controles de calidad necesarios para asegurar que no había habido ningún problema durante el análisis y que los resultados obtenidos se ajustaban a la realidad de estos tumores.

Analisis de datos

Análisis bioinformático

- Identificación de genes relevantes

La metodología usada se basa en la selección de características de tipo embebido. Se empleó nueve modelos de clasificación diferentes disponibles en Scikit-learn. Estos modelos comparten la existencia de métricas internas de selección de características, clave a la hora de seleccionar los marcadores con mayor capacidad predictiva. La lista

de modelos es: Lasso, Ridge, Gradient Boosting, Random Forest, ExtraTrees, LogisticRegression, SGDC, Passive Aggressive Classifier y SVR.

El conjunto de datos es balanceado con respecto al evento recaída (26 pacientes con recaída frente a 28 pacientes sin recaída), de modo que no hizo falta emplear procedimientos especiales para corregir este sesgo. Como preprocesamiento de la información, se procedió a un escalado y normalizado estándar de la expresión de los marcadores.

Para cada modelo se efectuaron simulaciones con una partición del 70% de datos del conjunto de entrenamiento y un 30% del conjunto de validación. En cada modelo, se entrenó con todos los marcadores (730 genes) y se identificó en cada simulación, mediante métricas propias de cada modelo, los 20 genes de más relevancia.

Estos genes eran los seleccionados como importantes en cada modelo y simulación.

Para consolidar los resultados, se efectuó mil simulaciones por cada modelo. A partir de ello se generó un vector de relevancia que asignaba a cada gen un valor de relevancia correspondiente a la proporción de veces que dicho marcador se identificó como importante. Adicionalmente se definió el rendimiento de cada modelo como la mediana del área bajo la curva (AUC), en las mil simulaciones realizadas en cada modelo. Esta métrica es adecuada para evaluar problemas de clasificación equilibrados.

Distintos modelos asignan relevancias diferentes a distintos genes. Ciertas discrepancias entre ellos son esperables, pero es importante definir un método objetivo que permita juzgar cuáles son los genes más relevantes de forma global, independientemente del modelo seleccionado. En este trabajo, se utilizó dos criterios: uno de unanimidad (el gen es relevante en todos los modelos) y otro de voto mayoritario (al menos la mitad de los modelos lo han considerado relevante).

Para establecer un orden de importancia entre los diferentes marcadores seleccionados se recurrió al método de elevación de umbral progresivo; es decir, se fue considerando los conjuntos de n genes más relevantes de cada modelo y se fue incrementando n. Para cada valor de n se determina qué genes están presentes en todos los conjuntos (criterio de unanimidad) o en la mayoría de ellos (criterio de voto mayoritario).

- Mejoras en predicción

Una vez identificados los genes clave, se comprobó finalmente que la capacidad de predicción de los modelos mejoraba, restringidos a los marcadores más relevantes. Para ello se comparó los valores de AUC en estimación de recaídas para el conjunto de todos

los genes y para secuencias de número creciente de los genes marcados como relevantes.

Resultados

Análisis de expresión génica diferencial de los 730 genes

Inicialmente se llevó a cabo un análisis de expresión diferencial usando los 730 genes en los 54 pacientes incluidos, usando la herramienta nSolver versión 5.

Con los resultados obtenidos con el análisis de expresión génica de las 54 muestras tumorales mediante el panel PanCancer (28 muestras de pacientes sin recaída y 26 muestras de pacientes con recaída tumoral) se realizó el análisis de expresión diferencial obteniendo, como vemos en la imagen tipo volcán de la *figura 3*, una probable expresión diferencial pero sin diferencias significativas de expresión en ninguno de los genes analizados, probablemente por el fold change no suficientemente amplio. Prácticamente todos los genes se encontraban en el rango entre doble infraexpresión y doble sobreexpresión al comparar la expresión de dichos genes entre ambos grupos (recaída y no recaída).

En la *tabla 1* se representan los genes con expresión diferencial más próximo a la significación estadística ($p < 0.05$).

Tabla 1. Significación estadística de los genes más representativos.

Gen	Fold change	p adj
AKT3	-0.62	0.064
BAIAP3	-0.90	0.064
ZBTB16	-0.90	0.064
RASGRP1	-0.92	0.075
KIT	-1.25	0.075
IL13RA2	0.96	0.075
ITGA3	-0.73	0.075
ACVR1B	-0.54	0.075
TGFB3	-0.89	0.075
HHEX	-0.80	0.075
CCNA1	0.97	0.075
NR4A1	-0.68	0.075
PLA2G10	0.64	0.075
FGF7	-1	0.075
GADD45A	0.64	0.075
MMP9	-1.31	0.082
TGFB2	-1.10	0.087
RRAS2	0.72	0.087
CDC7	0.33	0.095
CCNB1	0.61	0.095

Significación estadística de los genes con expresión diferencial más próxima a la significación ($p < 0.05$) (p adj) y fold change de cada uno de esos genes indicando el número de veces que cada gen está sobre/infraexpresado en el grupo con recaída respecto al grupo sin recaída (como vemos todos ellos entre -1 y +1).

Analisis bioinformático

Finalmente se llevó a cabo un análisis bioinformático usando distintas herramientas con el objetivo de disminuir el ruido ocasionado por el alto número de variables analizadas (730 genes). Se usó nueve modelos de clasificación (representados en la *tabla 2*), los cuales comparten métricas internas de selección de características, fundamental para seleccionar los genes con mayor capacidad predictiva.

Se realizaron simulaciones con cada modelo con una participación del 70% de datos del conjunto de entrenamiento y un 30% del conjunto de validación. En cada modelo se entrenó con los 730 genes y, se seleccionó en cada una de las mil simulaciones realizadas con cada modelo, los 20 genes con más relevancia de cada simulación. La mediana del rendimiento de cada modelo en las mil simulaciones realizadas con cada uno de ellos se representó como el área bajo la curva (AUC) (valor mediano de la AUC sobre las 1000 iteraciones), estando la mayoría en torno a 0.6 (valorando cada modelo individualmente). Los datos de entrada han sido meramente normalizados. Ninguna selección de marcadores ni procesado previo adicional ha tenido lugar. Los modelos han sido ejecutados con los valores defecto de los hiperparámetros de la librería scikit-learn.

Una guía adecuada para clasificar la precisión indicada por la AUC sería similar al sistema académico de puntuación:

- 0.90-1.00: clasificador excelente (A)
- 0.80-0.90: clasificador bueno (B)
- 0.70-0.80: clasificador aceptable (C)
- 0.60-0.70: clasificador mediocre (D)
- 0.50-0.60: clasificador deficiente (E)

En la *tabla 2* se representan los entrenamientos iniciales con los resultados para cada modelo empleado (730 marcadores empleados).

Tabla 2. Resultados de AUC con los entrenamientos iniciales en cada modelo.

MODELO	AUC
Lasso	0.667
Ridge	0.639
Gradient Boosting	0.530
Random Forest	0.586
ExtraTrees	0.598
LogisticRegression	0.607
SGDC	0.607
PassiveAgressiveClassifier	0.614
SVR	0.653
Resultado referencia (mediana AUC)	0.607

En la *figura 4* vemos el número de genes seleccionados, con cada uno de los criterios, en función del tamaño umbral empleado. Es decir, según el criterio de unanimidad (genes relevantes en todos los modelos, representado en rojo) el número de genes seleccionados variaba según el número de genes empleados en la simulación (a mayor número de genes empleados (tamaño muestral) más genes seleccionados). Igual ocurre cuando empleamos el criterio de voto mayoritario (genes relevantes al menos en la mitad de los modelos usados, representado en azul), donde a más genes empleados/estudiados más genes seleccionados/más genes se repiten como destacados en al menos la mitad de los modelos usados.

La secuencia de genes en orden de importancia (voto mayoritario) resulta ser: MPL, COL24A1, NR4A1, FOSL1, CREB5, FANCB, LAMB4, ZBTB16, CALML3.

A continuación se realizó las mismas simulaciones con un conjunto creciente de genes seleccionados, y se comparó los resultados con los mismos modelos utilizados sobre el conjunto total de genes (*tabla 3*).

Tabla 3. Simulaciones con un número creciente de genes seleccionados y resultados de AUC.

GENES SELECCIONADOS	MEDIANA AUC (AUC referencia 0.607)	MEJORA (%)
MPL	0.519	-14%
MPL, Col24a1	0.639	5%
MPL, Col24a1, NR4A1	0.735	21%
MPL, Col24a1, NR4A1, FOSL1	0.708	17%
MPL, Col24a1, NR4A1, FOSL1, CREB5	0.708	17%
MPL, Col24a1, NR4A1, FOSL1, CREB5, FANCB	0.736	21%
MPL, Col24a1, NR4A1, FOSL1, CREB5, FANCB, LAMB4	0.757	25%
MPL, Col24a1, NR4A1, FOSL1, CREB5, FANCB, LAMB4, ZBTB16	0.763	26%
MPL, COL24A1, NR4A1, FOSL1, CREB5, FANCB, LAMB4, ZBTB16, CALML3	0.790	30%
Mayor cantidad de genes	< 0.790	< 30%

Pensando en identificar una dependencia entre la expresión de los genes referencia y la recaída, habría que insistir en que no es una cuestión fácilmente abordable. Muchos de los modelos empleados son no-lineales y los comportamientos que predicen son bastante más complejos que una relación lineal basada en marcadores independientes. Además el análisis es bastante dependiente del split. Sin embargo, y dado que el modelo Ridge (lineal) presenta un buen comportamiento dentro de los modelos usados sobre el conjunto limitado de genes de referencia (mediana AUC > 0.80) parece relevante notar la dependencia que apunta sobre los genes y su relación con la recaída. Según sus coeficientes, tendríamos que la recaída es más probable cuando se da la siguiente relación para cada gen.

- MPL: sobreexpresión
- Col24a1: infraexpresión
- NR4A1: sobreexpresión

- FOSL1: sobreexpresión
- CREB5: sobreexpresión
- FANCB: sobreexpresión
- LAMB4: infraexpresión
- ZBTB16: infraexpresado
- CALML3: sobreexpresión

A la vista de los resultados concluimos que la expresión de los genes de la lista ofrece una cierta capacidad predictiva en términos de recaída, con una mejora de hasta un 30% de la mediana de la métrica AUC en el caso de la utilización de nueve marcadores referencia (MPL, COL24A1, NR4A1, FOSL1, CREB5, FANCB, LAMB4, ZBTB16, CALML3), aumentando el AUC de 0.5 a 0.8. Interpretamos que la razón por la que los modelos funcionan mejor seleccionando estos genes frente a los 730 totales se debe a la eliminación del ruido que suponen los restantes. Es interesante observar que, si aumentamos el número de marcadores referencia, los resultados comienzan a empeorar; posiblemente debido a la utilización de otros marcadores que suponen la introducción de más ruido que capacidad predictiva.

En resumen, se ha identificado en pacientes con TCGNS estadio I tratados exclusivamente con orquiectomía una firma génica de nueve genes (MPL, Col24a1, NR4A1, FOSL1, CREB5, FANCB, LAMB4, ZBTB16, CALML3) que permite seleccionar aquellos pacientes con una alta probabilidad de recaída (AUC 0.8).

REIVINDICACIONES

1.- *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía.

2.- Un método *in vitro* de obtención de datos útiles para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía que comprende medir en una muestra biológica aislada del individuo el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*.

3.- El método *in vitro* de obtención de datos útiles para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía según la reivindicación anterior, que además comprende comparar las cantidades obtenidas con una cantidad de referencia.

4.- Un método *in vitro* para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía que comprende medir en una muestra biológica aislada el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16,* comparar las cantidades obtenidas con una cantidad de referencia, e incluir al individuo en el grupo de individuos con una mayor probabilidad de recaída, cuando

a) MLP está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

b) *LAMB4* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

c) *FANCB* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

d) *NR4A1* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

e) *CREB5* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

f) *CALML3* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

g) *FOSL1* está sobreexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia;

h) *COL24A1* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia; y

i) *ZBTB16* está infraexpresado al menos 1,3 veces, preferiblemente al menos 1,4 veces, y aún más preferiblemente al menos 1,5 veces con respecto a una muestra de referencia.

5.- El método según cualquiera de las reivindicaciones 2-4, donde la muestra biológica es una muestra de tejido.

6.- El método según cualquiera de las reivindicaciones 2-5, donde la determinación del producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* se realiza mediante:

a) un procedimiento de análisis de la expresión génica, como por ejemplo qRT-PCR, microarrays de ADN, nCounter, RNA-Seq, FISH, microarrays de tejidos o una transferencia Northern; y/o

b) un inmunoensayo o inmunohistoquímica.

7.- El método según cualquiera de las reivindicaciones 2-5, donde la determinación del producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16* se realiza mediante nCounter.

8.- Un kit o dispositivo, que comprende los elementos y/o reactivos necesarios para cuantificar el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*.

9.- El kit o dispositivo según la reivindicación anterior, que comprende cebadores, sondas y/o anticuerpos capaces de cuantificar el producto de expresión de *MPL, LAMB4, FANCB, NR4A1, CREB5, CALML3, FOSL1, COL24A1 y ZBTB16*, y donde:

- los cebadores o primers son secuencias de polinucleótidos de entre 10 y 30 pares de bases, más preferiblemente de entre 15 y 25 pares de bases, aún más preferiblemente de entre 18 y 22 pares de bases, y aún mucho más preferiblemente de alrededor de 20 pares de bases, que presentan una identidad de al menos un 80%, más preferiblemente de al menos un 90%, aún más preferiblemente de al menos un 95%, aún mucho más preferiblemente de al menos un 98%, y particularmente de un 100%, con un fragmento de la secuencia complementaria a la SEQ ID NO: 2; SEQ ID NO: 4; SEQ ID NO: 6; SEQ ID NO: 8; SEQ ID NO:10; SEQ ID NO: 12; SEQ ID NO:14; SEQ ID NO: 16 y/o SEQ ID NO: 18;
- las sondas son secuencias de polinucleótidos de entre 30 y 1100 pares de bases, más preferiblemente de entre 100 y 1000 pares de bases, y aún más preferiblemente de entre 150 y 500 pares de bases, que presentan una identidad de al menos un 80%, más 25 preferiblemente de al menos un 90%, aún más preferiblemente de al menos un 95%, aún mucho más preferiblemente de al menos un 98%, y particularmente de un 100%, con un fragmento de las secuencias complementarias a la SEQ ID NO: 2; SEQ ID NO: 4; SEQ ID NO: 6; SEQ ID NO: 8; SEQ ID NO:10; SEQ ID NO: 12; SEQ ID NO:14; SEQ ID NO: 16 y/o SEQ ID NO: 18 y
- los anticuerpos son capaces de unirse específicamente a una región formada por la secuencia aminoacídica SEQ ID NO: 1, SEQ ID NO: 3, SEQ ID NO: 5; SEQ ID NO: 7; SEQ ID NO: 9; SEQ ID NO: 11; SEQ ID NO: 13; SEQ ID NO: 15 y/o SEQ ID NO: 17.

10-. El uso de un kit o dispositivo de la invención según cualquiera de las reivindicaciones 8 o 9, para predecir o pronosticar el riesgo de recaída en un individuo con tumor de células germinales no seminoma estadio I tras orquiectomía.

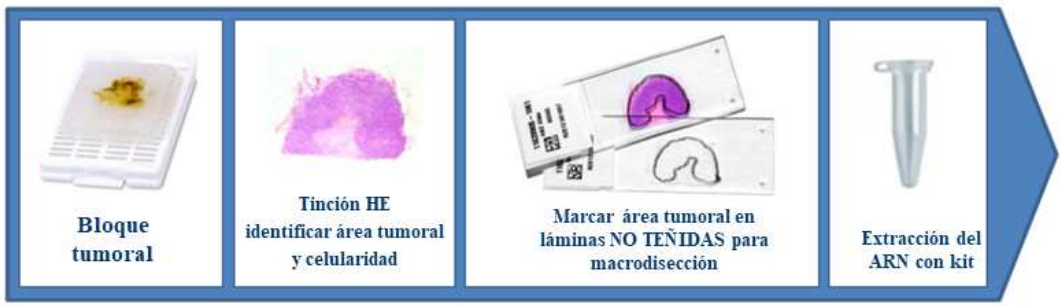
Bloque tumoral
Tinción HE identificar área tumoral y celularidad
Marcar área tumoral en láminas NO TEÑIDAS para macrodisección
Extracción del ARN con kit

Fig. 1


1. Hibridación
Sonda de captura
Sonda reporter
RNA diana
Hibrido: Sonda+RNA diana
2. Retirada exceso de sonda/purificación, alineamiento
3. DA: Contaje de moléculas
Code
Gene
Count
x
3
y
2
z
1

Fig. 2

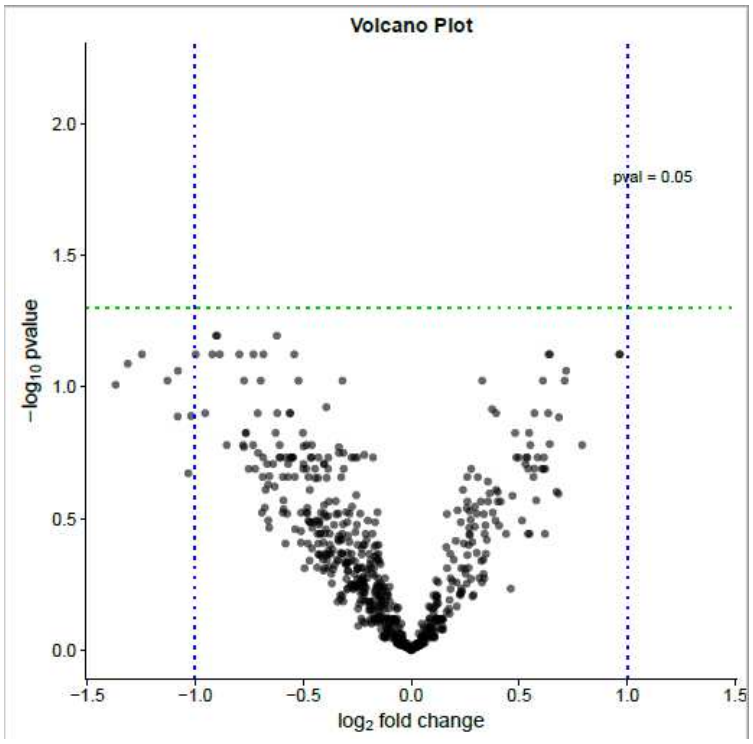
Volcano Plot
pval = 0.05
2.0
1.5
1.0
0.5
0.0
−log10 pvalue
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
log2 fold change

Fig. 3

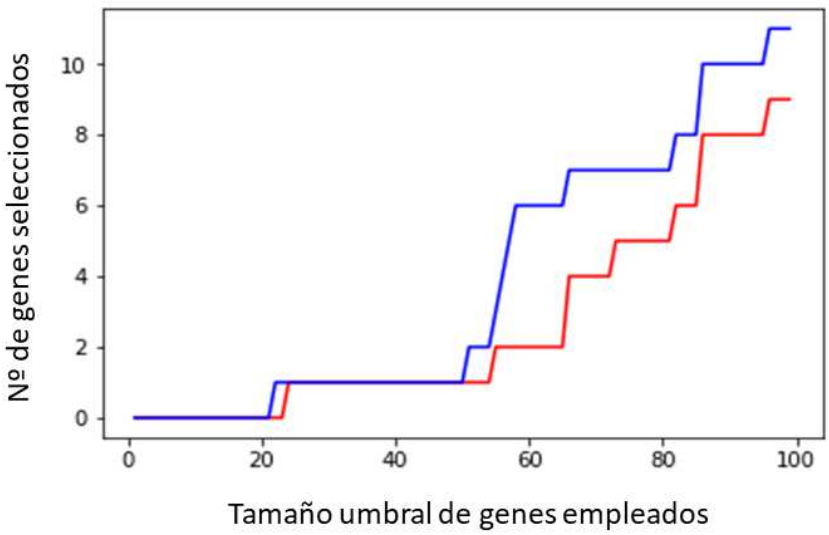
Nº de genes seleccionados
0
2
4
6
8
10
0
20
40
60
80
100
Tamaño umbral de genes empleados

Fig. 4