(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 358 910**

(51) Int. Cl.:
C12Q 1/68 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Número de solicitud europea: **05747218 .5**
(96) Fecha de presentación : **02.06.2005**
(97) Número de publicación de la solicitud: **1766056**
(97) Fecha de publicación de la solicitud: **28.03.2007**

(54) Título: **Oligonucleótidos para diagnósticos de cáncer de mama.**

(30) Prioridad: **02.06.2004 GB 0412301**

(45) Fecha de publicación de la mención BOPI:
16.05.2011

(45) Fecha de la publicación del folleto de la patente:
16.05.2011

(73) Titular/es: **DIAGENIC AS.**
Ostensjoveien 158
0661 Oslo, NO
Elizabeth Louise Jones

(72) Inventor/es: **Sharma, Praveen y**
Lönneborg, Anders

(74) Agente: **Elzaburu Márquez, Alberto**

DESCRIPCIÓN

El presente invento se refiere a sondas oligonucleotídicas para uso en la evaluación de los niveles de transcritos génicos en una célula, que pueden ser utilizadas en técnicas analíticas, particularmente en técnicas diagnósticas. Convenientemente, las sondas se proporcionan en forma de kit. Se pueden utilizar diferentes conjuntos de sondas en técnicas para preparar patrones de expresión génica e identificar, diagnosticar o controlar cánceres de mama.

La identificación de métodos rápidos y sencillos para el análisis de muestras para, por ejemplo, aplicaciones diagnósticas sigue siendo el objetivo de muchos investigadores. Los usuarios finales buscan métodos que sean eficaces en cuanto al costo, produzcan resultados estadísticamente significativos y puedan ser rutinariamente implementados sin necesidad de individuos muy especializados.

El análisis de la expresión génica dentro de las células ha sido utilizado para obtener información sobre el estado de esas células e, importantemente, el estado del individuo del que proceden las células. La expresión relativa de diversos genes en una célula ha sido identificada como un reflejo de un estado particular de un organismo. Por ejemplo, se sabe que las células cancerosas presentan una expresión alterada de diversas proteínas y, por lo tanto, se pueden usar los transcritos o las proteínas expresadas como marcadores de ese estado morboso.

De esta manera, se puede analizar el tejido de una biopsia en cuanto a la presencia de esos marcadores y se pueden identificar células procedentes del sitio de la enfermedad en otros tejidos o fluidos del organismo por la presencia de los marcadores. Además, se pueden liberar productos de la expresión alterada en la corriente sanguínea y se pueden analizar esos productos. Además, las células que han entrado en contacto con las células de la enfermedad pueden verse afectadas por su contacto directo con esas células para dar lugar a una expresión génica alterada, y su expresión o los productos de su expresión pueden ser similarmente analizados.

Sin embargo, hay ciertas limitaciones con estos métodos. Por ejemplo, el uso de marcadores tumorales específicos para identificar el cáncer adolece de una diversidad de defectos, tales como falta de especificidad o sensibilidad, asociación del marcador con estados morbosos además de con el tipo específico de cáncer, y dificultad para la detección en individuos asintomáticos.

Más recientemente, además del análisis de una o dos proteínas o transcritos marcadores, se han analizado patrones de expresión génica. La mayoría del trabajo, que acarrea el análisis de la expresión génica a gran escala con implicaciones en el diagnóstico de la enfermedad, ha requerido muestras clínicas procedentes de células o tejidos enfermos. Por ejemplo, en varias publicaciones recientes, que demuestran que se pueden usar datos de expresión génica para distinguir entre tipos de cáncer similares, se han utilizado muestras clínicas procedentes de células o tejidos enfermos (Alon et al., 1999, PNAS 96, páginas 6745-6750; Golub et al., 1999, Science 286, páginas 531-537; Alizadeh et al., 2000, Nature 403, páginas 503-511; Bittner et al., 2000, Nature 406, páginas 536-540).

Sin embargo, estos métodos se han basado en el análisis de una muestra que contiene células enfermas o productos de esas células o células que han entrado en contacto con células de la enfermedad. El análisis de dichas muestras depende del conocimiento de la presencia de una enfermedad y su localización, lo que puede ser difícil en pacientes asintomáticos. Además, no siempre se pueden tomar muestras del sitio de la enfermedad; por ejemplo, en enfermedades del cerebro.

En un hallazgo de gran significación, los presentes inventores identificaron el potencial previamente no explotado de todas las células de un organismo para obtener información relativa al estado del organismo del cual procedían las células. En el Documento WO98/49342 se describe el análisis de la expresión génica de células alejadas del sitio de la enfermedad, por ejemplo, de sangre periférica recogida lejos de un sitio canceroso. En el Documento WO 2004/046382 se describen sondas específicas para el diagnóstico del cáncer de mama y la enfermedad de Alzheimer y se discuten protocolos para identificar otras sondas apropiadas para ese fin y para diagnosticar otras enfermedades.

Este hallazgo se basa en la premisa de que las diferentes partes del cuerpo de un organismo existen en interacción dinámica entre sí. Cuando una enfermedad afecta a una parte del cuerpo, otras partes del cuerpo resultan también afectadas. La interacción es el resultado de un amplio espectro de señales bioquímicas que se liberan de la zona enferma, para afectar a otras zonas del cuerpo. Aunque la naturaleza de los cambios bioquímicos y fisiológicos inducidos por las señales liberadas puede variar en las diferentes partes del cuerpo, los cambios pueden ser medidos a nivel de la expresión génica y ser utilizados con fines diagnósticos.

El estado fisiológico de una célula en un organismo viene determinado por el patrón con que se expresan genes en ella. El patrón depende de los estímulos biológicos internos y externos a los que está expuesta dicha célula, y cualquier cambio en el grado o la naturaleza de estos estímulos puede conducir a un cambio en el patrón con que se expresan los diferentes genes en la célula. Existe el conocimiento creciente de que, analizando los cambios sistémicos en patrones de expresión génica en células de muestras biológicas, es posible obtener información sobre el tipo y la naturaleza de los estímulos biológicos que están actuando sobre ellas. De este modo, por ejemplo, controlando la expresión de un gran número de genes en células de una muestra de ensayo, es posible determinar si sus genes se expresan con un patrón característico de una enfermedad, un estado o una fase particular de los mismos. Por lo tanto la medición de cambios en las actividades génicas de células, por ejemplo, de tejidos o fluidos corporales, está surgiendo como una potente herramienta para el diagnóstico de enfermedades.

Dichos métodos presentan varias ventajas. A menudo, la obtención de muestras clínicas de ciertas zonas del cuerpo que está enfermo puede ser difícil y puede implicar invasiones indeseables del cuerpo; por ejemplo, a menudo se utiliza la biopsia para obtener muestras de cáncer. En ciertos casos, tal como en la enfermedad de Alzheimer, la muestra del cerebro enfermo sólo se puede obtener post mórtem. Además, las muestras tisulares que se obtienen son a menudo heterogéneas y pueden contener una mezcla de células tanto enfermas como no enfermas, lo que hace que el análisis de los datos sobre expresión génica generados sea complejo y difícil.

Se ha sugerido que una colección de tejidos tumorales que parecen ser patogenéticamente homogéneos con respecto a aspectos morfológicos del tumor bien puede ser muy heterogénea a nivel molecular (Alizadeh, 2000, *supra*) y, en realidad, podría contener tumores que representaran enfermedades esencialmente diferentes (Alizadeh, 2000, *supra*; Golub, 1999, *supra*). Con el fin de identificar una enfermedad, un estado o una fase de los mismos, es muy deseable cualquier método que no requiera que las muestras clínicas procedan directamente de células o tejidos enfermos ya que, de una región fácilmente accesible del organismo, se pueden obtener muestras clínicas que representan una mezcla homogénea de tipos celulares.

Hemos identificado ahora una familia de secuencias que permite la obtención de un conjunto de sondas de sorprendente utilidad para identificar el cáncer de mama. De este modo, describimos ahora familias de genes cuya expresión está alterada en las células de muestras sanguíneas procedentes de pacientes con cáncer de mama, que se pueden usar para generar sondas para uso en métodos para identificar, diagnosticar o controlar el cáncer de mama o fases del mismo.

En el trabajo que condujo a este invento, los inventores examinaron el nivel de expresión de un gran número de genes en pacientes con cáncer con respecto al nivel en pacientes normales. No sólo se halló que había un gran número de genes que presentaban una expresión alterada sino que, además, se halló que aquellos que presentaban una expresión alterada caían dentro de familias discretas de genes en virtud de su función. Como tales, estos genes proporcionan una colección a partir de la cual se pueden generar las correspondientes sondas que pueden ser utilizadas colectivamente para generar una huella dactilar de la expresión de estos genes en un individuo. Puesto que la expresión de estos genes está alterada en el individuo con cáncer y, por consiguiente, puede ser considerada informativa de ese estado, la huella dactilar generada a partir de la colección de sondas es indicativa de la enfermedad con respecto al estado normal.

Las familias de genes que han sido identificadas por expresarse diferencialmente en pacientes con cáncer pueden resumirse del modo siguiente.

(i) genes que codifican proteínas implicadas en la síntesis y/o estabilidad de proteínas;

(ii) genes que codifican proteínas implicadas en la regulación de la defensa y/o la remodelación de cromatina.

La familia (i) incluye:

(a) genes que codifican proteínas ribosómicas y proteínas de activación ribosómica (es decir, proteínas que comprenden componentes de proteínas ribosómicas o implicadas en la modificación de su función y de las que se ha hallado que están infrarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen las proteínas ribosómicas L1-L56, L7A, L10A, L13A, L18A, L23A, L27A, L35A, L36A, L37A, P0, P1, P2, S2-S29, S31, S33-S36, S3A, S15A, S18A, S18B, S18C, S27A, 63, 115 (y seudogenes), proteína cinasas ribosómicas (por ejemplo, la cinasa S6), ribonucleasas, la proteína del supuesto dominio S1 ligante de RNA, factores de inicio de la traducción eucarióticos y la proteína G ligante de nucleótidos de guanina;

(b) genes que codifican factores de iniciación e inhibición de la traducción (es decir, proteínas implicadas en la traducción de mRNA en un producto proteico y de las que se ha hallado que están infrarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen factores de elongación de la traducción eucarióticos, tRNA sintetasas, proteínas ligantes de RNA, proteínas ligantes de elementos de poliadenilación, tirosina fosfatasas, factores de inicio de la traducción eucarióticos, y factores de transcripción de las RNA polimerasas I, III;

(c) genes que codifican otros agentes moduladores de la transcripción o la traducción, tales como la proteína ligante de tipo ciclina D y la proteína ligante de nucleótidos de guanina.

La familia (ii) incluye:

(a) genes que codifican proteínas relacionadas con la respuesta inmune (es decir, proteínas que están suprarreguladas en respuesta a una estimulación inmune y que incluyen proteínas suprarreguladas en respuesta a una inflamación o en la generación de una respuesta inflamatoria, y de las que se ha hallado que están suprarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen el receptor de células T y componentes asociados, tales como, por ejemplo, proteína cinasas, diversas citocinas, incluyendo las interleucinas y sus receptores (tales como IL-1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 17, 18, 20, 22 y 24), el factor de necrosis tumoral y su receptor y su superfamilia (por ejemplo, los miembros 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14 y 15 de la superfamilia del TNF), factores reguladores de interferones, oncostatina M, factor inhibidor de la leucemia, la familia de ligandos y receptores de quimiocinas (por ejemplo, los números 1-28), componentes del complemento, factores estimulados por interferones, tales como factores de transcripción,

MHC (por ejemplo, HLA) de clase I o clase II (o componentes relacionados) (por ejemplo, DQ, DR, DO, DP y DM alfa o beta), proteínas de adhesión (por ejemplo, CD1A, CD1C, CD1D, CD3Z, 6, 8, 11, 14, 18, 24, 27, 28, 29, 40, 44, 50, 54, 59, 74, 79B, 80, 81, 83, 86, 96 e ICAM), el factor nuclear del potenciador del gen del polipéptido kappa en células B, la proteína básica de mielina, catepsina, el receptor de tipo peaje, subunidades de proteosomas, ferritina, proteína cinasas o fosfatasas así como sus activadores e inhibidores, el receptor de tipo inmunoglobulina de leucocitos, componentes de inmunoglobulinas, tales como, por ejemplo, la cadena pesada o fragmentos Fc de, por ejemplo, IgG, IgE o IgA o su superfamilia, defensina, oxitocina, la proteína S100 ligante de calcio, lectina y su receptor y superfamilia, leptina, fosfolipasa y factores de crecimiento (tales como el factor de crecimiento de células endoteliales y la eritropoyetina);

(b) genes que codifican proteínas inducidas por TNF (es decir, proteínas que son inducidas en un individuo en respuesta a la exposición a TNF y de las que se ha hallado que están suprarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen la proteína 8 inducida por TNF alfa, integrina, el inhibidor del potenciador del gen del polipéptido ligero kappa en células B, los factores 2 y 5 asociados con TNF, el factor nuclear del potenciador del gen del polipéptido ligero kappa en células B, MAP cinasas, proteína cinasa C, cinasa ubicua, cadherina, caspasa, ciclina D1, superóxido dismutasa e interleucinas;

(c) genes que codifican proteínas inducidas por hipoxia (es decir, proteínas que resultan inducidas cuando el individuo o parte de él está en un estado de hipoxia y de las que se ha hallado que están suprarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen sestrina, la proteína p300 ligante de E1A, endotelina, la proteína relacionada con la ataxia telangiectasia y Rad3, hexocinasa 2, la tirosina cinasa TEK, el factor de fragmentación de DNA, caspasa, activador de plasminógeno, factor 1 inducible por hipoxia, y glucosa fosfato isomerasa;

(d) genes que codifican proteínas de estrés oxidativo (es decir, proteínas que resultan inducidas en un individuo o parte de él bajo un estrés oxidativo y de las que se ha hallado que están suprarreguladas en pacientes con cáncer). Estas proteínas codificadas incluyen superóxido dismutasa, glutatión sintetasa, catalasa, lactoperoxidasa, peroxidasa tiroidea, mieloperoxidasa, peroxidasa de eosinófilo, resistencia 1 a la oxidación, peroxirredoxina, citocromo P450, receptor supresor ("scavenger"), paraoxonasa, glutatión reductasa, NAD(P)H deshidrogenasa, glutatión S-transferasa, catenina, glutarredoxina, proteínas de choque térmico (tales como factores de transcripción de choque térmico), proteína cinasas activadas por mitógenos, enolasa, tiorredoxina reductasa y peroxirredoxina;

(e) genes que codifican proteínas implicadas en la remodelación de cromatina (es decir, proteínas que sirven de instrumento para mantener o modificar la estructura de la cromatina y pueden ser esenciales para la regulación génica). Estas proteínas codificadas incluyen proteínas de reemplazo de histonas, tal como, por ejemplo, la familia H3.3A o H3.3B.

Las apropiadas secuencias génicas que caen dentro de las familias anteriormente descritas pueden ser identificadas mediante una búsqueda en bases de datos apropiadas utilizando, como palabra clave, el nombre de la familia, por ejemplo, "respuesta inmune" en las bases de datos de genes o proteínas del "National Centre for Biotechnology Information" de Noruega. Para la confirmación de la utilidad de dichas secuencias génicas para el desarrollo de oligonucleótidos para los ensayos aquí descritos, se puede evaluar la expresión de una secuencia génica particular en un paciente de ensayo con cáncer frente a un paciente normal. Una variación de expresión por encima o debajo de los niveles testigo es indicativa de la utilidad de la secuencia para la obtención de sondas.

Generalmente, los genes que codifican las familias (i) anteriores están infrarregulados en los pacientes con cáncer frente a los pacientes normales, y, en el caso de las familias (ii), los genes de codificación están suprarregulados.

Se especula con que, en pacientes con cáncer, la expresión disminuida sistemática de los genes implicados en la producción de ribosomas y el control de la traducción puede indicar que células sanguíneas están respondiendo a un nuevo estado en esos pacientes al disminuir la velocidad de síntesis proteica, lo que puede ser una adaptación celular a un ambiente de bajo oxígeno y déficit energético. Esto viene respaldado por la observación de que genes implicados en la defensa contra especies oxigenadas reactivas (ROS; del inglés, reactive oxygen species), tales como MnSOD y ferritina, están suprarregulados en las muestras cancerosas. Un bajo nivel de eritropoyetina puede explicar los bajos niveles de oxígeno en los pacientes con cáncer. También se cree que la activación del TNF es una vía para los cambios en las familias de genes anteriormente descritas ya que se sabe que el TNF suprarregula la expresión de, por ejemplo, ferritina, defensina, MnSOD y calgranulina B. El TNF también inhibe la producción de EPO, lo que puede causar un estado de bajo oxígeno en el ambiente sanguíneo. Se sabe que la hipoxia induce niveles de TNF. Estos cambios pueden ser desencadenados por factores angiogénicos que entran en la corriente sanguínea. Aunque sin pretender un respaldo teórico, en la Figura 1 se muestra la hipótesis que se oculta tras los efectos anteriormente descritos.

De este modo, el invento proporciona un conjunto de sondas oligonucleotídicas, como se proporciona en las reivindicaciones, que corresponden a genes de una célula cuya expresión está afectada según un patrón característico del cáncer de mama, en que dichos genes son sistémicamente afectados por dicho cáncer. Preferiblemente, dichos genes se expresan constitutivamente de forma moderada o elevada. Preferiblemente, los genes se expresan moderada o elevadamente en las células de la muestra pero no en células de la enfermedad ni en células que han entrado en contacto con dicha células de la enfermedad.

Dichas sondas, particularmente cuando se aíslan de células alejadas del sitio de la enfermedad, no dependen del desarrollo de la enfermedad hasta niveles clínicamente reconocibles y permiten la detección del cáncer de mama muy pronto

tras el inicio de dicho cáncer, incluso años antes de que aparezcan otros síntomas subjetivos u objetivos.

Como aquí se usa, genes "sistémicamente" afectados se refiere a genes cuya expresión está afectada en el organismo sin contacto directo con una célula de la enfermedad ni con el sitio de la enfermedad, y las células bajo investigación no son células de la enfermedad.

"Contacto", como aquí se hace referencia, se refiere a células que entran en íntima proximidad entre sí de modo que se puede observar el efecto directo de una célula sobre la otra, por ejemplo, una respuesta inmune, en que estas respuestas no son mediadas por moléculas secundarias liberadas por la primera célula a larga distancia para afectar a la segunda célula. Preferiblemente, el contacto se refiere a un contacto físico o al contacto que es tan próximo como es estéricamente posible; convenientemente, las células que entran en contacto se encuentran en el mismo volumen unitario, por ejemplo, en 1 cm^3.

Una "célula de la enfermedad" es una célula que manifiesta cambios fenotípicos y está presente en el sitio de la enfermedad en algún momento de su vida, tal como, por ejemplo, una célula tumoral en el sitio tumoral o que se ha diseminado desde el tumor, o una célula cerebral en el caso de cáncer del cerebro.

Genes expresados "moderada o elevadamente" se refiere a los que están presentes en células en reposo con un número de copias superior a 30-100 copias/célula (suponiendo un valor medio de $3x10^5$ moléculas de mRNA en una célula).

Se proporcionan aquí sondas específicas que tienen las propiedades anteriormente descritas.

De este modo, en un aspecto, el presente invento proporciona un conjunto de sondas oligonucleotídicas, en que dicho conjunto consiste en menos de 500 oligonucleótidos y comprende todos los oligonucleótidos enumerados en la Tabla 2 ó 3, en que cada oligonucleótido puede ser sustituido por un oligonucleótido que es una parte de dicho oligonucleótido de la Tabla 2 ó 3 o por un oligonucleótido con una secuencia complementaria.

Se describen aquí sondas oligonucleotídicas seleccionadas de entre los oligonucleótidos enumerados en la Tabla 2, 3 ó 4 (por ejemplo, de la Tabla 2) o derivados de una secuencia descrita en la Tabla 2, 3 ó 4, o una secuencia complementaria de la misma. Dichos oligonucleótidos derivados incluyen oligonucleótidos derivados de los genes que corresponden a las secuencias proporcionadas en esas tablas, por ejemplo, los genes expuestos en las Tablas 2, 5 y 6 (véanse los números de acceso), o las secuencias complementarias de las mismas. El uso del conjunto de sondas del invento en productos y métodos del invento forma aspectos adicionales del invento.

Como aquí se hace referencia, un "oligonucleótido" es una molécula de ácido nucleico que tiene al menos 6 monómeros, es decir, nucleótidos o formas modificadas de los mismos, en la estructura polímera. La molécula de ácido nucleico puede ser DNA, RNA o ácido nucleico peptídico (PNA; del inglés, peptide nucleic acid) o híbridos de los mismos o versiones modificadas de los mismos, por ejemplo, formas químicamente modificadas, por ejemplo, ácido nucleico cerrado (LNA; del inglés, locked nucleic acid), por metilación o compuestas de bases modificadas o artificiales durante la síntesis, con tal de que conserven su capacidad para unirse a secuencias complementarias. Dichos oligonucleótidos se usan para sondar secuencias diana y, por lo tanto, a ellos se hace también referencia aquí como sondas oligonucleotídicas o, simplemente, sondas.

Un oligonucleótido que corresponde a una secuencia génica de la familia (i) o (ii) se refiere a un oligonucleótido que corresponde a toda, o parte de, dicha secuencia génica o su transcrito. Cuando se utiliza una parte de la secuencia génica, satisface los requisitos de las sondas oligonucleotídicas que aquí se describen en cuanto a, por ejemplo, longitud y función. Preferiblemente, dichas partes tienen el tamaño descrito más adelante. A dicho oligonucleótido se hace referencia más adelante como oligonucleótido primario. Un oligonucleótido derivado se refiere a un oligonucleótido que es parte del oligonucleótido primario pero satisface los requisitos para sondas que aquí se describen.

Preferiblemente, las sondas oligonucleotídicas que forman dicho conjunto tienen una longitud de al menos 15 bases para permitir la unión de las moléculas diana. Especialmente, dichas sondas oligonucleotídicas tienen preferiblemente una longitud de 20 a 200 bases, por ejemplo, de 30 a 150 bases, preferiblemente una longitud de 50-100 bases.

Como aquí se hace referencia, la expresión "secuencias complementarias" se refiere a secuencias con bases complementarias consecutivas (es decir, T:A, G:C), secuencias complementarias que, por lo tanto, pueden unirse entre sí a través de su complementariedad.

Los oligonucleótidos de los conjuntos del invento son como se describen en la Tabla 2 ó 3 o una parte de un oligonucleótido descrito en la Tabla 2 ó 3, o su secuencia complementaria. Dicho conjunto puede comprender además una o más sondas oligonucleotídicas enumeradas en la Tabla 4, o una parte de un oligonucleótido descrito en la Tabla 4, o una secuencia complementaria del mismo.

Como se describe, un "conjunto" se refiere a una colección de sondas oligonucleotídicas únicas (es decir, que tienen una secuencia distinta) y consiste en menos de 500 sondas oligonucleotídicas, por ejemplo, preferiblemente de 10 a 500, tal como, por ejemplo, de 10 a 100, 200 ó 300, especialmente preferiblemente de 20 a 100, por ejemplo, de 30 a 100 sondas. En ciertos casos se pueden utilizar menos de 10 sondas, tal como, por ejemplo, de 2 a 9 sondas, por ejemplo, de 5 a 9 sondas.

Se apreciará que aumentar el número de sondas evitará la posibilidad de un mal análisis, tal como, por ejemplo, una falta de diagnóstico por comparación con otras enfermedades que podrían alterar similarmente la expresión de los genes particulares en cuestión. También pueden estar presentes otras sondas oligonucleotídicas no descritas aquí, particularmente si ayudan al uso último del conjunto de sondas oligonucleotídicas. Sin embargo, preferiblemente, dicho conjunto sólo consiste en los oligonucleótidos aquí descritos o en un subconjunto de los mismos de acuerdo con las reivindicaciones.

En cada conjunto pueden estar presentes múltiples copias de cada sonda oligonucleotídica única, por ejemplo, 10 o más copias, pero sólo constituyen una sola sonda.

Un conjunto de sondas oligonucleotídicas, que pueden estar preferiblemente inmovilizadas sobre un soporte sólido o tienen medios para dicha inmovilización, comprende las sondas oligonucleotídicas que se describieron antes. Como se mencionó anteriormente, estas sondas deben ser únicas y tener secuencias diferentes. Sin embargo, una vez dicho esto, se pueden usar dos sondas distintas que reconozcan el mismo gen pero reflejen diferentes procesos de corte y empalme, con tal de que todos los oligonucleótidos enumerados en la Tabla 2 ó 3 o su reemplazo, como aquí se describe, estén presentes en dicho conjunto. Sin embargo, se prefieren las sondas oligonucleotídicas que son complementarias de, y se unen a, genes distintos.

Como aquí se describe, un oligonucleótido "funcionalmente equivalente" o derivado se refiere a un oligonucleótido que permite identificar el mismo gen que un oligonucleótido de una secuencia de las familias de secuencias génicas aquí descritas; es decir, se puede unir a la misma molécula de mRNA (o DNA) transcrita a partir de un gen (molécula de ácido nucleico diana) que el oligonucleótido primario o el oligonucleótido derivado (o su secuencia complementaria). Preferiblemente, dicho oligonucleótido derivado o funcionalmente equivalente es una parte de una secuencia génica como la definida en la Tabla 2, 5 ó 6, o la secuencia complementaria de la misma. Preferiblemente, dicho oligonucleótido funcionalmente equivalente es capaz de reconocer, es decir, unirse a, el mismo producto de corte y empalme que un oligonucleótido primario o un oligonucleótido derivado. Preferiblemente, dicha molécula de mRNA es la molécula de mRNA de longitud completa que corresponde al oligonucleótido primario o el oligonucleótido derivado.

Como aquí se hace referencia, "capaz de unirse" o "unirse" se refiere a la capacidad para hibridarse bajo condiciones descritas más adelante. De acuerdo con el invento, los oligonucleótidos derivados son aquellos que son una parte de un oligonucleótido enumerado en la Tabla 2, 3 ó 4.

Como aquí se utiliza, una "parte" se refiere a un tramo de al menos 5, por ejemplo, al menos 10 ó 20 bases, tal como de 5 a 100, por ejemplo, de 10 a 50 o de 15 a 30 bases, de dicho oligonucleótido primario.

Como se describió anteriormente, dicho conjunto de sondas oligonucleotídicas puede ser convenientemente inmovilizado sobre uno o más soportes sólidos. A dichos soportes sólidos se fija una sola copia o, preferiblemente, múltiples copias de cada sonda única; por ejemplo, están presentes 10 o más, por ejemplo, al menos 100, copias de cada sonda única.

Se pueden asociar una o más sondas oligonucleotídicas únicas con soportes sólidos separados que forman juntos un conjunto de sondas inmovilizadas sobre múltiples soportes sólidos; por ejemplo, se pueden inmovilizar una o más sondas únicas sobre múltiples glóbulos, membranas, filtros, biochips, etc., que forman juntos un conjunto de sondas, que forman juntas módulos del kit descrito más adelante. Convenientemente, los soportes sólidos de los diferentes módulos están físicamente asociados, aunque las señales asociadas con cada sonda (generadas como se describe más adelante) deben ser separadamente determinables. Alternativamente, se pueden inmovilizar las sondas sobre porciones discretas del mismo soporte sólido; por ejemplo, se puede inmovilizar cada sonda oligonucleotídica única, por ejemplo, en múltiples copias, sobre una porción o región distinta y discreta de un solo filtro o membrana para, por ejemplo, generar una red.

También se puede utilizar una combinación de dichas técnicas; por ejemplo, se pueden utilizar diversos soportes sólidos cada uno de los cuales inmoviliza varias sondas únicas.

La expresión "soporte sólido" significará cualquier material sólido capaz de unirse a oligonucleótidos mediante puentes hidrófobos, iónicos o covalentes.

Como aquí se utiliza, "inmovilización" se refiere a la asociación reversible o irreversible de las sondas con dicho soporte sólido en virtud de dicha unión. Si es reversible, las sondas permanecen asociadas con el soporte sólido durante el tiempo suficiente para que se lleven a cabo los métodos del invento.

Numerosos soportes sólidos adecuados como componentes de inmovilización de acuerdo con el invento son bien conocidos en la técnica y ampliamente descritos en la bibliografía, y, hablando en términos generales, el soporte sólido puede ser cualquiera de los soportes o matrices bien conocidos que actualmente se utilizan o proponen mucho para la inmovilización, separación, etc., en procedimientos químicos o bioquímicos. Dichos materiales incluyen, pero no se limitan a, cualquier polímero orgánico sintético tal como poliestireno, poli(cloruro de vinilo) o polietileno; o nitrocelulosa y acetato de celulosa; o superficies activadas con tosilo; o vidrio o nailon o cualquier superficie que porta un grupo adecuado para la copulación covalente de ácidos nucleicos. Los componentes de inmovilización pueden tomar la forma de partículas, láminas, geles, filtros, membranas, tiras de microfibra, tubos o placas, fibras o capilares, hechos de, por

ejemplo, un material polímero tal como, por ejemplo, agarosa, celulosa, alginato, teflón, látex o poliestireno o glóbulos magnéticos. Se prefieren los soportes sólidos que permiten la presentación de una red, preferiblemente en una sola dimensión, tales como, por ejemplo, láminas, filtros, membranas, placas o biochips.

La fijación de las moléculas de ácido nucleico al soporte sólido puede ser llevada a cabo directa o indirectamente. Por ejemplo, si se utiliza un filtro, la fijación puede ser llevada a cabo mediante entrecruzamiento inducido con luz UV. Alternativamente, la fijación puede ser llevada indirectamente a cabo mediante el uso de un componente de fijación contenido en las sondas oligonucleotídicas y/o el soporte sólido. De esta manera, por ejemplo, se puede usar una pareja de miembros ligantes por afinidad, tal como avidina, estreptavidina o biotina, DNA o una proteína ligante de DNA (por ejemplo, la proteína represora lac I o la secuencia operadora lac a la cual se une), anticuerpos (que pueden ser monoclonales o policlonales), fragmentos de anticuerpo o los epítopos o haptenos de anticuerpos. En estos casos, un miembro de la pareja ligante se fija a (o es una parte inherente de) el soporte sólido y el otro miembro se fija a (o es una parte inherente de) las moléculas de ácido nucleico.

Como aquí se utiliza, una "pareja ligante por afinidad" se refiere a dos componentes que se reconocen y unen específicamente entre sí (es decir, con preferencia a la unión a otras moléculas). Dichas parejas ligantes forman un complejo cuando se unen.

La fijación de grupos funcionales apropiados al soporte sólido puede ser llevada a cabo mediante métodos bien conocidos en la técnica, los cuales incluyen, por ejemplo, la fijación por medio de grupos hidroxilo, carboxilo, aldehído o amino que se pueden proporcionar por tratamiento del soporte sólido para proporcionar revestimientos superficiales adecuados. Los soportes sólidos que presentan componentes apropiados para la fijación del miembro ligante pueden ser producidos mediante métodos rutinarios conocidos en la técnica.

La fijación de grupos funcionales apropiados a las sondas oligonucleotídicas aquí descritas puede ser llevada a cabo por ligación o ser introducida durante la síntesis o la multiplicación, utilizando, por ejemplo, cebadores que portan un componente apropiado, tal como biotina o una secuencia particular para captura.

Convenientemente, el conjunto de sondas anteriormente descrito se proporciona en forma de kit.

De este modo, visto desde otro aspecto, el presente invento proporciona un kit que comprende un conjunto de sondas oligonucleotídicas como las anteriormente descritas, inmovilizadas sobre uno o más soportes sólidos.

Preferiblemente, dichas sondas están inmovilizadas sobre un solo soporte sólido, y cada sonda única está fijada a una región diferente de dicho soporte sólido. Sin embargo, cuando se fijan a múltiples soportes sólidos, dichos múltiples soportes sólidos forman los módulos que componen el kit. Especialmente, dicho soporte sólido es preferiblemente una lámina, filtro, membrana, placa o biochip.

Opcionalmente, el kit puede contener también información relativa a las señales generadas por muestras normales o enfermas (como se discute más adelante con mayor detalle en relación con el uso de los kits), materiales para estandarización, por ejemplo, mRNA o cDNA de muestras normales y/o enfermas con fines comparativos, etiquetas para incorporación a cDNA, adaptadores para introducir secuencias de ácido nucleico con fines de multiplicación, cebadores para multiplicación y/o enzimas apropiadas, tampones y disoluciones. Opcionalmente, dicho kit puede contener también una información añadida que describe cómo se debería llevar el método del invento a cabo, que proporciona opcionalmente gráficos, datos o software estándares para la interpretación de los resultados obtenidos cuando se lleva el invento a cabo.

El uso de dichos kits para preparar un patrón diagnóstico estándar de transcritos génicos como el descrito más adelante forma un aspecto más del invento.

El conjunto de sondas como el aquí descrito tiene diversos usos. Sin embargo, se usa principalmente para evaluar el estado de la expresión génica de una célula de ensayo para obtener información relativa al organismo del cual procede dicha célula. De este modo, las sondas son útiles para diagnosticar, identificar o controlar el cáncer de mama en un organismo.

De este modo, en un aspecto más, el invento proporciona el uso de un conjunto de sondas oligonucleotídicas o un kit como el anteriormente descrito para determinar el patrón de expresión génica de una célula, patrón que refleja el nivel de expresión génica de los genes a los que se unen dichas sondas oligonucleotídicas, para el diagnóstico del cáncer de mama, que comprende al menos las operaciones de:

a) aislar mRNA de dicha célula, que opcionalmente puede ser inversamente transcrito a cDNA;

b) hibridar el mRNA o cDNA de la operación (a) con un conjunto de sondas oligonucleotídicas o un kit como el aquí descrito; y

c) evaluar la cantidad de mRNA o cDNA que se hibrida con cada una de dichas sondas, para producir dicho patrón.

El mRNA y el cDNA, como a ellos se hace referencia en este método, y los métodos posteriores abarcan derivados o

copias de dichas moléculas, por ejemplo, copias de dichas moléculas tales como las producidas por multiplicación o la preparación de cadenas complementarias, pero que conservan la identidad de la secuencia de mRNA, es decir, se hibridarían con el transcrito directo (o su secuencia complementaria) en virtud de una complementariedad exacta, o una identidad de secuencias, a lo largo de al menos una región de dicha molécula. Se apreciará que no existirá complementariedad a lo largo de la región entera cuando se han usado técnicas que pueden truncar el transcrito o introducir nuevas secuencias, por ejemplo, mediante multiplicación de cebadores. Por conveniencia, dicho mRNA o cDNA se multiplica preferiblemente antes de la operación b). Como los oligonucleótidos aquí descritos, dichas moléculas pueden ser modificadas utilizando, por ejemplo, bases artificiales durante la síntesis, con tal de que permanezca la complementariedad. Dichas moléculas pueden llevar también componentes adicionales tales como medios para señalización o inmovilización.

Más adelante se describen con mayor detalle las diversas operaciones implicadas en el método para preparar dicho patrón.

Como aquí se utiliza, "expresión génica" se refiere a la transcripción de un gen particular para producir un producto de mRNA específico (es decir, un producto de corte y empalme particular). El nivel de expresión génica puede ser determinado evaluando el nivel de moléculas de mRNA transcritas o moléculas de cDNA inversamente transcritas a partir de las moléculas de mRNA o productos derivados de esas moléculas, por ejemplo, por multiplicación.

El "patrón" creado mediante esta técnica se refiere a una información que, por ejemplo, se puede representar en forma tabular o gráfica y contiene información acerca de la señal asociada con dos o más oligonucleótidos. Preferiblemente, dicho patrón se expresa como un agrupamiento de números relativos al nivel de expresión asociado con cada sonda.

Preferiblemente, dicho patrón se establece utilizando el siguiente modelo lineal:

$$y = Xb + f \qquad \text{Ecuación 1}$$

en que X es la matriz de los datos de expresión génica e *y* es la variable respuesta, *b* es el vector coeficiente de regresión y *f* es el vector residual estimado. Aunque se pueden utilizar muchos métodos diferentes para establecer la relación proporcionada en la ecuación 1, especialmente se utiliza preferiblemente el método de Regresión por Mínimos Cuadrados Parciales (PLSR; del inglés, Partial Least Squares Regression) para establecer la relación de la ecuación 1.

Se usan así las sondas para generar un patrón que refleja la expresión génica de una célula en el momento de su aislamiento. El patrón de expresión es característico de las circunstancias bajo las cuales se encuentra esa célula y depende de las influencias a que ha estado expuesta la célula. De este modo, se puede preparar una huella dactilar o patrón estándar característico de transcritos génicos (patrón de sondas estándar) para células de un individuo con cáncer de mama y se puede utilizar para la comparación con patrones de transcritos de células de ensayo. Esto tiene aplicaciones evidentes en cuanto a diagnosticar, controlar o identificar si un organismo está aquejado de cáncer de mama.

Se prepara el patrón estándar determinando el grado de unión de las sondas con el mRNA total (o el cDNA o producto relacionado) procedente de células de una muestra de uno o más organismos con el cáncer. Esto refleja el nivel de transcritos que están presentes, que corresponden a cada sonda única. Se valora la cantidad de material de ácido nucleico que se une a las diferentes sondas, y esta información forma el patrón estándar de transcritos génicos de ese cáncer. Cada uno de dichos patrones estándares es característico del cáncer.

Por lo tanto, en un aspecto más, el presente invento proporciona un método para preparar un patrón estándar de transcritos génicos característico del cáncer de mama en un organismo, que comprende al menos las operaciones de:

a) aislar mRNA de las células de una muestra de uno o más organismos que tienen cáncer de mama, que opcionalmente puede ser inversamente transcrito a cDNA;

b) hibridar el mRNA o cDNA de la operación (a) con un conjunto de oligonucleótidos como los anteriormente descritos, específicos para dicho cáncer de mama en un organismo y una muestra del mismo que corresponde al organismo y una muestra del mismo bajo investigación; y

c) evaluar la cantidad de mRNA o cDNA que se hibrida con cada una de dichas sondas para producir un patrón característico que refleja el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en la muestra con cáncer de mama.

Por conveniencia, dichos oligonucleótidos están preferiblemente inmovilizados sobre uno o más soportes sólidos.

Se puede acumular el patrón estándar para cáncer de mama en bases de datos y hacerlo disponible para los laboratorios que lo soliciten.

Como aquí se hace referencia, muestras y organismos de la "enfermedad" o muestras y organismos del "cáncer" se refieren a organismos (o muestras de los mismos) con una proliferación celular anormal, por ejemplo, en una masa sólida tal como un tumor. Se sabe que dichos organismos tienen o presentan el cáncer de mama bajo estudio.

Como aquí se utiliza, "normal" se refiere a organismos o muestras que se usan con fines comparativos. Preferiblemente, estos son "normales" en el sentido de que no presentan ninguna señal de, o no se cree que tengan, enfermedad o estado alguno que afecte a la expresión génica, particularmente en relación con el cáncer de mama para el que se van a usar como estándar normal.

Como aquí se utiliza, una "muestra" se refiere a cualquier material obtenido del organismo, por ejemplo, un ser humano o un animal no humano bajo investigación, que contiene células e incluye tejidos, fluidos corporales o desechos corporales o, en el caso de organismos procarióticos, el propio organismo. Los "fluidos corporales" incluyen sangre, saliva, líquido de la médula espinal, semen y linfa. El "desecho corporal" incluye orina, materia expectorada (pacientes pulmonares), heces, etc. Las "muestras tisulares" incluyen tejido obtenido mediante biopsia, mediante intervenciones quirúrgicas o mediante otros medios, por ejemplo, de placenta. Sin embargo, las muestras que se examinan proceden preferiblemente de zonas del organismo aparentemente no afectadas por el cáncer. Las células de dichas muestras no son células de la enfermedad, es decir, células cancerosas, no han estado en contacto con dichas células de la enfermedad y no proceden del sitio del cáncer. Se considera que el "sitio de la enfermedad" es esa zona del cuerpo que manifiesta la enfermedad de una manera que permite su determinación objetiva, tal como, por ejemplo, un tumor. De este modo, por ejemplo, se puede utilizar sangre periférica para el diagnóstico del cáncer de mama, y la sangre no requiere la presencia de células malignas o diseminadas procedentes del cáncer.

Sin embargo, se apreciará que el método para preparar el patrón de transcripción estándar y otros métodos del invento son también aplicables para uso sobre partes vivas de organismos eucarióticos, tales como líneas celulares y cultivos y explantes de órganos.

Como aquí se utiliza, la referencia a una muestra "correspondiente", etc., se refiere preferiblemente a células del mismo tejido, fluido corporal o desecho corporal, pero también incluye células de tejido, fluido corporal o desecho corporal que son suficientemente similares para el fin de preparar el patrón estándar o de ensayo. Cuando se utiliza en referencia a genes "que corresponden" a las sondas, esto se refiere a genes que están relacionados por secuencia (que puede ser complementaria) con las sondas aunque las sondas pueden reflejar diferentes productos de expresión por corte y empalme.

Como aquí se utiliza, "evaluar" se refiere a la evaluación tanto cuantitativa como cualitativa, que puede ser determinada en términos absolutos o relativos.

El invento puede ser llevado a la práctica del modo siguiente.

Para preparar un patrón de transcritos estándar para el cáncer de mama, se extrae mRNA de muestra de las células de tejidos, fluido corporal o desecho corporal de un individuo u organismo enfermo de acuerdo con técnicas conocidas [véase, por ejemplo, Sambrook et al. (1989), Molecular Cloning: A Laboratory Manual, 2ª edición, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, EE.UU.].

Debido a las dificultades para trabajar con RNA, el RNA es preferiblemente transcrito inversamente en esta fase para formar cDNA de primera cadena. Sin embargo, la clonación del cDNA o la selección a partir de, o utilizando, un banco de cDNA no es necesaria en éste ni otros métodos del invento. Preferiblemente, se sintetizan las cadenas complementarias de los cDNAs de primera cadena, es decir, cDNAs de segunda cadena, pero esto dependerá de qué cadenas relativas están presentes en las sondas oligonucleotídicas. Sin embargo, alternativamente, el RNA puede ser utilizado directamente sin transcripción inversa y, si así se requiere, puede ser etiquetado.

Preferiblemente, las cadenas de cDNA son multiplicadas mediante técnicas de multiplicación conocidas, tal como la reacción en cadena de la polimerasa (PCR; del inglés, polymerase chain reaction) mediante el uso de cebadores apropiados. Alternativamente, las cadenas de cDNA pueden ser clonadas con un vector y usadas para transformar bacterias tales como *E. coli*, que pueden ser luego cultivadas para multiplicar las moléculas de ácido nucleico. Cuando no se conocen las secuencias de los cDNAs, se pueden dirigir cebadores a regiones de las moléculas de ácido nucleico que se han introducido. De este modo, por ejemplo, se pueden ligar adaptadores a las moléculas de cDNA y se pueden dirigir cebadores a estas porciones para la multiplicación de las moléculas de cDNA. Alternativamente, en el caso de muestras eucarióticas, se pueden aprovechar la cola de poliA y la caperuza del RNA para preparar cebadores apropiados.

Para producir la huella dactilar o patrón diagnóstico estándar de transcritos génicos para el cáncer de mama, se utilizan las sondas oligonucleotídicas anteriormente descritas para sondar mRNA o cDNA de la muestra enferma para producir una señal para hibridación con cada especie de sonda oligonucleotídica particular, es decir, cada sonda única. Si se desea, también se puede preparar un patrón testigo estándar de transcritos génicos utilizando mRNA o cDNA de una muestra normal. De este modo, se pone mRNA o cDNA en contacto con la sonda oligonucleotídica bajo condiciones apropiadas que permitan la hibridación.

Cuando se sondan múltiples muestras, esto puede ser llevado consecutivamente a cabo usando las mismas sondas, por ejemplo, sobre uno o más soportes sólidos, es decir, sobre módulos del kit de sondas, o hibridando simultáneamente con las sondas correspondientes, por ejemplo, los módulos de un kit de sondas correspondientes.

Para identificar cuándo se produce la hibridación y obtener una indicación del número de transcritos/moléculas de cDNA

que se llegan a unir a las sondas oligonucleotídicas, es necesario identificar una señal producida cuando los transcritos (o moléculas relacionadas) se hibridan (por ejemplo, por detección de moléculas de ácido nucleico de doble cadena o detección del número de moléculas que se llegan a unir, después de la separación de las moléculas no unidas mediante, por ejemplo, lavado).

Con objeto de obtener una señal, uno cualquiera o los dos componentes que se hibridan (es decir, la sonda y el transcrito) portan o forman un medio de señalización o una parte del mismo. Este "medio de señalización" es cualquier componente que permite la detección directa o indirecta mediante la generación o presencia de una señal. La señal puede ser cualquier característica física detectable, tal como la conferida por emisión de radiación, propiedades de dispersión o absorción, propiedades magnéticas, u otras propiedades físicas tales como propiedades de carga, tamaño o unión de moléculas existentes (por ejemplo, etiquetas) o moléculas que se pueden generar (por ejemplo, emisión de gases, etc.). Se prefieren las técnicas que permiten la multiplicación de la señal, por ejemplo, que producen múltiples procesos de señal a partir de un solo sitio ligante activo, tal como, por ejemplo, mediante la acción catalítica de enzimas para producir múltiples productos detectables.

Convenientemente, el medio de señalización puede ser una etiqueta que proporcione por sí misma una señal detectable. Convenientemente esto puede ser llevado a cabo mediante el uso de un agente radiactivo u otra etiqueta que se puede incorporar durante la producción del cDNA, la preparación de cadenas de cDNA complementarias o durante la multiplicación del mRNA/cDNA diana, o se puede añadir directamente a moléculas de ácido nucleico diana.

Las etiquetas apropiadas son aquéllas que permiten directa o indirectamente la detección o medición de la presencia de los transcritos/cDNA. Dichas etiquetas incluyen, por ejemplo, radioetiquetas, etiquetas químicas, tales como, por ejemplo, agentes cromóforos o fluoróforos (por ejemplo, colorantes tales como fluoresceína y rodamina), y reactivos de alta densidad electrónica tales como ferritina, hemocianina y oro coloidal. Alternativamente, la etiqueta puede ser una enzima, tal como, por ejemplo, peroxidasa o fosfatasa alcalina, visualizándose la presencia de la enzima por su interacción con una sustancia adecuada, tal como, por ejemplo, un sustrato. La etiqueta también puede formar parte de una pareja de señalización en que el otro miembro de la pareja se encuentra en, o muy próximo a, la sonda oligonucleotídica a la que se une el transcrito/cDNA; por ejemplo, se puede utilizar un compuesto fluorescente y un sustrato sofocador de la fluorescencia. También se puede disponer una etiqueta en una sustancia diferente, tal como un anticuerpo, que reconozca un componente peptídico fijado a los transcritos/cDNA, por ejemplo, fijado a una base utilizada durante la síntesis o la multiplicación.

Se puede obtener una señal mediante la introducción de una etiqueta antes, durante o después de la operación de hibridación. Alternativamente, la presencia de transcritos de hibridación puede ser identificada mediante otras propiedades físicas, tal como su absorbancia, en cuyo caso, el medio de señalización es el propio complejo.

Luego se evalúa la cantidad de señal asociada con cada sonda oligonucleotídica. La evaluación puede ser cuantitativa o cualitativa y se puede basar en la unión de una sola especie de transcrito (o cDNA relacionado u otros productos) a cada sonda, o la unión de múltiples especies de transcrito a múltiples copias de cada sonda única. Se apreciará que los resultados cuantitativos proporcionarán más información sobre la huella dactilar de transcritos del cáncer de mama que se compila. Estos datos se pueden expresar como valores absolutos (en el caso de macrorredes) o se pueden determinar con respecto a una referencia o estándar particular, tal como, por ejemplo, una muestra testigo normal.

Además, se apreciará que el patrón diagnóstico estándar de transcritos génicos puede ser preparado usando una o más muestras de la enfermedad (y muestras normales, si se utilizan) para llevar a cabo la operación de hibridación con objeto de obtener patrones no sesgados hacia unas variaciones de expresión génica particulares del individuo.

El uso del conjunto de sondas anteriormente descrito para preparar patrones estándares y los patrones diagnósticos estándares de transcritos génicos así producidos con el fin de identificación, diagnóstico o control del cáncer de mama en un organismo particular forma un aspecto más del invento.

Una vez que se ha determinado un patrón o huella dactilar diagnóstica estándar para el cáncer de mama utilizando el conjunto de sondas oligonucleotídicas, se puede utilizar esta información para identificar la presencia, ausencia o grado de cáncer de mama en un organismo o individuo de ensayo diferente.

Para examinar el patrón de expresión génica de una muestra de ensayo, se obtiene de un paciente o del organismo que se estudia una muestra de ensayo de tejido, fluido corporal o desecho corporal que contiene células, correspondiente a la muestra utilizada para la preparación del patrón estándar. Luego se prepara un patrón de transcritos génicos de ensayo del modo anteriormente descrito para el patrón estándar.

Por lo tanto, en un aspecto más, el presente invento proporciona un método para preparar un patrón de transcritos génicos de ensayo, que comprende al menos las operaciones de:

a) aislar mRNA de las células de una muestra de dicho organismo de ensayo, que opcionalmente puede ser inversamente transcrito a cDNA;

b) hibridar el mRNA o cDNA de la operación (a) con un conjunto de oligonucleótidos como los anteriormente descritos, específicos para dicho cáncer de mama en un organismo y una muestra del mismo que corresponde al orga-

nismo y una muestra del mismo bajo investigación; y

c) evaluar la cantidad de mRNA o cDNA que se hibrida con cada una de dichas sondas para producir dicho patrón que refleja el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en dicha muestra de ensayo.

Este patrón de ensayo puede ser luego comparado con uno o más patrones estándares para evaluar si la muestra contiene células que tienen cáncer de mama.

De este modo, visto desde otro aspecto, el presente invento proporciona un método para diagnosticar o identificar o controlar el cáncer de mama en un organismo, que comprende las operaciones de:

a) aislar mRNA de las células de una muestra de dicho organismo, que opcionalmente puede ser inversamente transcrito a cDNA;

b) hibridar el mRNA o cDNA de la operación (a) con un conjunto de oligonucleótidos como los anteriormente descritos, específicos para dicho cáncer de mama en un organismo y una muestra del mismo que corresponde al organismo y una muestra del mismo bajo investigación;

c) evaluar la cantidad de mRNA o cDNA que se hibrida con cada una de dichas sondas para producir un patrón característico que refleje el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en dicha muestra; y

d) comparar dicho patrón con un patrón diagnóstico estándar preparado de acuerdo con el método del invento utilizando una muestra de un organismo que corresponde al organismo y la muestra bajo investigación para determinar la presencia de cáncer de mama en el organismo bajo investigación.

El método hasta la operación c) inclusive es la preparación de un patrón de ensayo como se describió anteriormente.

Como aquí se hace referencia, "diagnóstico" se refiere a la determinación de la presencia o existencia de cáncer de mama en un organismo. "Controlar" se refiere a establecer el grado del cáncer de mama, particularmente cuando se sabe que un individuo padece cáncer de mama, para, por ejemplo, controlar los efectos del tratamiento o el desarrollo del cáncer de mama para, por ejemplo, determinar la idoneidad de un tratamiento o proporcionar un pronóstico.

Se puede determinar la presencia de cáncer de mama determinando el grado de correlación entre los patrones de las muestras estándar y de ensayo. Esto toma necesariamente en consideración el intervalo de los valores que se obtienen para muestras normales y enfermas. Aunque esto se puede establecer obteniendo desviaciones estándares para diversas muestras representativas que se unen a las sondas para desarrollar el estándar, se apreciará que pueden bastar muestras individuales para generar el patrón estándar para identificar el cáncer de mama si la muestra de ensayo presenta una correlación lo suficientemente cercana con la estándar. Convenientemente, se puede pronosticar la presencia, ausencia o grado del cáncer de mama en una muestra de ensayo al insertar los datos relativos al nivel de expresión de sondas informativas de la muestra de ensayo en el patrón diagnóstico estándar de sondas establecido de acuerdo con la ecuación 1.

Los datos generados al utilizar los métodos anteriormente mencionados se pueden analizar usando diversas técnicas, desde la representación visual más básica (por ejemplo, en relación con la intensidad) hasta la manipulación de datos más complejos para identificar patrones subyacentes que reflejen la interrelación del nivel de expresión de cada gen al que se unen las diversas sondas, que puede ser cuantificado y expresado matemáticamente. Convenientemente, los datos brutos así generados pueden ser manipulados mediante los métodos estadísticos y de procesamiento de datos descritos más adelante, particularmente normalizando y estandarizando los datos y ajustando los datos a un modelo de clasificación para determinar si dichos datos de ensayo reflejan el patrón de cáncer de mama.

Los métodos aquí descritos pueden ser usados para identificar, controlar o diagnosticar el cáncer de mama o su progreso, del que son informativas las sondas oligonucleotídicas. Como aquí se describe, sondas "informativas" son aquéllas que reflejan genes que tienen una expresión alterada en el cáncer de mama. Puede que las sondas utilizadas en los conjuntos del invento no sean suficientemente informativas con fines diagnósticos cuando se utilizan solas, pero son informativas cuando se utilizan como una de varias sondas para proporcionar un patrón característico, por ejemplo, en un conjunto como el anteriormente descrito y de acuerdo con el invento.

Preferiblemente, dichas sondas corresponden a genes que resultan sistémicamente afectados por el cáncer de mama. Especialmente, dichos genes, de los que proceden los transcritos que se unen a las sondas utilizadas en conjuntos del invento, se expresan preferiblemente de forma moderada o elevada. La ventaja de usar sondas dirigidas a genes expresados moderada o elevadamente es que se requieren muestras clínicas más pequeñas para generar el necesario conjunto de datos de expresión génica, por ejemplo, muestras sanguíneas de menos de 1 ml.

Además, se ha hallado que dichos genes que ya se están transcribiendo activamente tienden a ser más propensos a resultar influidos, de un modo positivo o negativo, por nuevos estímulos. Además, puesto que ya se están produciendo transcritos en niveles que son generalmente detectables, pequeños cambios en esos niveles son fácilmente detectables

ya que, por ejemplo, no es necesario que se alcance un cierto umbral detectable.

Las sondas aquí descritas pueden ser usadas para el diagnóstico, identificación o control del cáncer de mama.

El método diagnóstico puede ser utilizado solo como una alternativa a otras técnicas diagnósticas o además de dichas técnicas. Por ejemplo, se pueden utilizar métodos del invento como una medida diagnóstica alternativa o aditiva al diagnóstico al usar técnicas de formación de imágenes tales como formación de imágenes por resonancia magnética (MRI; del inglés, magnetic resonance imaging), formación de imágenes por ultrasonidos, formación de imágenes por radionucleidos y formación de imágenes por rayos X en, por ejemplo, la identificación y/o el diagnóstico de tumores.

Los métodos del invento pueden ser llevados a cabo sobre células de organismos procarióticos o eucarióticos, que pueden ser cualesquier organismos eucarióticos tales como seres humanos, otros mamíferos y animales.

Los animales no humanos preferidos sobre los que se pueden llevar a cabo los métodos del invento incluyen, pero no se limitan a, mamíferos, particularmente primates, animales domésticos, animales de granja y animales de laboratorio. De este modo, los animales preferidos para el diagnóstico incluyen ratones, ratas, cobayas, gatos, perros, cerdos, vacas, cabras, ovejas y caballos. Particularmente, se diagnostica, identifica o controla preferiblemente el cáncer de mama de seres humanos.

Como se describió anteriormente, la muestra bajo estudio puede ser cualquier muestra conveniente que se puede obtener de un organismo. Sin embargo, como se mencionó anteriormente, la muestra se obtiene preferiblemente de un sitio alejado del sitio de la enfermedad y las células de dichas muestras no son células de la enfermedad, no han estado en contacto con dichas células y no proceden del sitio de la enfermedad. En tales casos, aunque preferiblemente ausentes, la muestra puede contener células que no satisfagan estos criterios. Sin embargo, puesto que las sondas de los conjuntos del invento tienen relación con transcritos cuya expresión está alterada en células que satisfacen estos criterios, las sondas se dirigen específicamente a detectar cambios en los niveles de transcritos de esas células incluso si están en presencia de otras células de fondo.

Se ha hallado que las células de dichas muestras muestran variaciones significativas e informativas en la expresión génica de un gran número de genes. De este modo, se puede hallar que la misma sonda (o varias sondas) es informativa en determinaciones relativas a dos o más cánceres, o fases de los mismos, en virtud del particular nivel de transcritos que se unen a esa sonda o de la interrelación del grado de unión a esa sonda con respecto a otras sondas. Como consecuencia, es posible utilizar un número relativamente pequeño de sondas para explorar múltiples cánceres. Esto tiene consecuencias en cuanto a la selección de sondas pero también en cuanto al uso de un solo conjunto de sondas para más de un diagnóstico.

Los métodos para generar patrones estándares y de ensayo y las técnicas diagnósticas dependen del uso de sondas oligonucleotídicas informativas para generar los datos de expresión génica. La metodología siguiente describe un método conveniente para identificar dichas sondas informativas.

Se pueden identificar sondas de diversos modos conocidos en la técnica previa, incluyendo mediante expresión diferencial y mediante resta de bancos (véase, por ejemplo, el Documento WO98/49342). Como se describió en el Documento WO 2004/046382 y como se describe más adelante, a la vista del elevado contenido de información de la mayor parte de los transcritos, como punto de partida uno también puede analizar simplemente un subconjunto aleatorio de especies de mRNA o cDNA que correspondan a la familia de secuencias aquí descrita y escoger las sondas más informativas de ese subconjunto. El método siguiente describe el uso de sondas oligonucleotídicas inmovilizadas (por ejemplo, las sondas del invento) a las que se une mRNA (o moléculas relacionadas) de diferentes muestras, para identificar qué sondas son las más informativas para identificar el cáncer de mama.

Las sondas inmovilizadas pueden proceder de diversos organismos relacionados o no relacionados; el único requisito es que las sondas inmovilizadas se unan específicamente a sus equivalentes homólogos de organismos de ensayo. Las sondas pueden también proceder de bases de datos públicas o comercialmente asequibles y estar inmovilizadas sobre soportes sólidos. Las sondas seleccionadas incluyen necesariamente el conjunto de sondas aquí descrito.

La longitud de las sondas inmovilizadas sobre el soporte sólido debe ser lo bastante grande para permitir la unión específica a las secuencias diana. Las sondas inmovilizadas pueden estar en forma de DNA, RNA o sus productos modificados o PNAs (ácidos nucleicos peptídicos). Preferiblemente, las sondas inmovilizadas deben unirse específicamente a sus equivalentes homólogos que representan genes elevada y moderadamente expresados de organismos de ensayo.

Se puede generar el patrón de expresión génica de células en muestras biológicas utilizando técnicas del campo técnico anterior, tal como una microrred o macrorred como la descrita más adelante, o utilizando métodos aquí descritos. Se han desarrollado ahora diversas tecnologías para controlar simultáneamente el nivel de expresión de un gran número de genes en muestras biológicas, tales como oligorredes de alta densidad (Lockhart et al., 1996, Nat. Biotech. 14, páginas 1675-1680), microrredes de cDNA (Schena et al., 1995, Science 270, páginas 467-470) y macrorredes de cDNA (E. Maier et al., 1994, Nucl. Acids Res. 22, páginas 3423-3424; Bernard et al., 1996, Nucl. Acids Res. 24, páginas 1435-1442).

En las oligorredes de alta densidad y las microrredes de cDNA, cientos y miles de oligonucleótidos sonda o cDNAs son

salpicados sobre portaobjetos de vidrio o membranas de nailon o sintetizados sobre biochips. Los mRNAs aislados de las muestras de ensayo y de referencia son etiquetados por transcripción inversa con un colorante fluorescente rojo o verde, mezclados, e hibridados con la microrred. Después de un lavado, los colorantes fluorescentes unidos son detectados mediante un láser, produciéndose dos imágenes, una para cada colorante. La relación resultante de las manchas rojas y verdes en las dos imágenes proporciona la información acerca de los cambios en los niveles de expresión de los genes en las muestras de ensayo y de referencia. Alternativamente, también se pueden llevar a cabo estudios con microrredes de un solo canal o de múltiples canales.

En la macrorred de cDNA, se salpican diferentes cDNAs sobre un soporte sólido, tal como membranas de nailon, en exceso en relación con la cantidad de mRNA de ensayo que se puede hibridar con cada mancha. El mRNA aislado de las muestras de ensayo es radiomarcado por transcripción inversa e hibridado con el cDNA sonda inmovilizado. Después de un lavado, se detectan y cuantifican las señales asociadas con etiquetas que se hibridan específicamente con el cDNA sonda inmovilizado. Los datos obtenidos en la macrorred contienen información acerca de los niveles relativos de transcritos presentes en las muestras de ensayo. Mientras que las macrorredes son sólo adecuadas para controlar la expresión de un número limitado de genes, las microrredes pueden ser utilizadas para controlar simultáneamente la expresión de varios miles de genes y, por lo tanto, son una elección preferida para estudios de expresión génica a gran escala.

Se ha utilizado una técnica de macrorredes para generar el conjunto de datos de expresión génica con objeto de ilustrar el método de identificación de sondas aquí descrito para la identificación de las sondas de los conjuntos del invento. Para este fin, se aísla mRNA de muestras de interés y se utiliza para preparar moléculas diana etiquetadas, tales como, por ejemplo, mRNA o cDNA, del modo anteriormente descrito. Las moléculas diana etiquetadas son luego hibridadas con sondas inmovilizadas sobre el soporte sólido. Para este fin, como se describió previamente, se pueden utilizar diversos soportes sólidos. Después de la hibridación, se separan las moléculas diana no unidas y se cuantifican las señales de las moléculas diana que se hibridan con sondas inmovilizadas. Si se lleva a cabo un radioetiquetado, se puede utilizar un sistema PhosphoImager para generar un archivo de imágenes que puede ser usado para generar un conjunto de datos brutos. Dependiendo de la naturaleza de la etiqueta escogida para etiquetar las moléculas diana, también se pueden utilizar otros instrumentos; por ejemplo, cuando se utiliza fluorescencia para el etiquetado, se puede usar un sistema FluoroImager para generar un archivo de imágenes procedente de las moléculas diana hibridadas.

Los datos brutos que corresponden a intensidad media, intensidad mediana o volumen de las señales de cada mancha pueden ser obtenidos del archivo de imágenes usando un software comercialmente asequible para el análisis de imágenes. Sin embargo, los datos obtenidos han de ser corregidos en cuanto a las señales de fondo y normalizados antes del análisis ya que diversos factores pueden afectar a la calidad y la cantidad de las señales de hibridación. Por ejemplo, variaciones en la calidad y cantidad del mRNA aislado de una muestra a otra, diferencias sutiles en la eficacia del etiquetado de las moléculas dianas durante cada reacción y variaciones en la cantidad de unión inespecífica entre diferentes macrorredes pueden contribuir al ruido del conjunto de datos obtenidos, que debe ser corregido antes del análisis.

La corrección del fondo puede ser llevada a cabo de diversas formas. Se puede utilizar la menor intensidad de píxel de una mancha para la resta del fondo o se puede utilizar para ese fin el valor medio o mediano de la línea de píxeles alrededor del contorno de las manchas. Uno también puede definir el área que representa la intensidad del fondo basándose en las señales generadas a partir de testigos negativos y utilizar la intensidad media de este área para la resta del fondo.

Los datos corregidos en cuanto al fondo pueden ser luego transformados para estabilizar la varianza de la estructura de datos y ser normalizados en cuanto a las diferencias en la intensidad de las sondas. Se han descrito diversas técnicas de transformación en la bibliografía, y de ellas se puede hallar una breve visión general en Cui, Kerr y Churchill, http://www.jax.org/research/churchill/research/expression/Cui-Transform.pdf). Se puede llevar a cabo la normalización dividiendo la intensidad de cada mancha por la intensidad colectiva, intensidad media o intensidad mediana de todas las manchas de una macrorred o de un grupo de manchas de una macrorred con objeto de obtener la intensidad relativa de las señales que se hibridan con sondas inmovilizadas en una macrorred. Se han descrito diversos métodos para normalizar datos de expresión génica (Richmond y Sommerville, 2000, Current Opin. Plant Biol. 3, páginas 108-116; Finkelstein et al., 2001, en "Methods of Microarray Data Analysis, Papers from CAMDA", redactado por Lin y Johnson, Kluwer Academica, páginas 57-68; Yang et al., 2001, en "Optical Technologies and Informatics", redactado por Bittner, Chen, Dorsel y Dougherty, Proceedings of SPIE, 4266, páginas 141-152; Dudoit et al., 2000, J. Am. Stat. Ass. 97, páginas 77-87; Alter et al., 2000, *supra*; Newton et al., 2001, J. Comp. Biol. 8, páginas 37-52). Generalmente, se calcula primero una función o factor de escalamiento para corregir el efecto de intensidad y se utiliza luego para normalizar las intensidades. También se ha sugerido el uso de testigos externos para una normalización mejorada.

Otro desafío importante encontrado en el análisis de la expresión génica a gran escala es el de la estandarización de los datos recogidos de experimentos llevados a cabo en diferentes momentos. Hemos observado que los datos de expresión génica de muestras adquiridas en el mismo experimento pueden ser eficazmente comparados después de la corrección del fondo y la normalización. Sin embargo, los datos de las muestras adquiridas en experimentos llevados a cabo en distintos momentos requieren otra estandarización antes del análisis. Esto se debe a que diferencias sutiles en parámetros experimentales entre diferentes experimentos, tales como, por ejemplo, diferencias en la calidad y la cantidad del mRNA extraído en diferentes momentos y diferencias en el tiempo utilizado para el etiquetado de moléculas diana, el tiempo de hibridación o el tiempo de exposición, pueden afectar a los valores medidos. Además, factores tales

como la naturaleza de la secuencia de los transcritos bajo investigación (su contenido de GC) y su cantidad relativa determinan cómo resultan afectados por sutiles variaciones en los procesos experimentales. Por ejemplo, determinan cómo cDNAs de primera cadena, que corresponden a un transcrito particular, son eficazmente transcritos y etiquetados durante la síntesis de la primera cadena, o cómo las moléculas diana etiquetadas correspondientes se unen eficazmente a sus secuencias complementarias durante la hibridación. Las diferencias de lote a lote en el proceso de impresión es también un factor importante para la variación en los datos de expresión generados.

Un fallo a la hora de abordar y rectificar apropiadamente estas influencias conduce a situaciones en que las diferencias entre las series experimentales pueden ensombrecer la principal información de interés contenida en el conjunto de datos de expresión génica, es decir, las diferencias en los datos combinados de las diferentes series experimentales. Por lo tanto, cuando sea necesario, los datos de expresión deben ser ajustados en cuanto a los lotes antes de su análisis.

El control de la expresión de un gran número de genes en varias muestras conduce a la generación de una gran cantidad de datos que son demasiado complejos para ser fácilmente interpretados. Ya se ha mostrado que diversas técnicas para el análisis supervisado y no supervisado de datos multivariantes son útiles para extraer una información biológica significativa de estos grandes conjuntos de datos. El análisis de grupos ("clusters") es con mucho la técnica más comúnmente usada para el análisis de la expresión génica y ha sido llevado a cabo para identificar genes que son regulados de una manera similar y/o identificar nuevas/desconocidas clases tumorales utilizando perfiles de expresión génica (Eisen et al., 1998, PNAS 95, páginas 14.863-14.868; Alizadeh et al., 2000, *supra*; Perou et al., 2000, Nature 406, páginas 747-752; Ross et al., 2000, Nature Genetics 24 (3), páginas 227-235; Herwig et al., 1999, Genome Res. 9, páginas 1093-1105; Tamayo et al., 1999, Science, PNAS, 96, páginas 2907-2912).

En el método de agrupamiento, se agrupan los genes en categorías funcionales (grupos) basándose en su perfil de expresión y satisfaciendo dos criterios: *homogeneidad* - los genes del mismo grupo tienen una expresión muy similar entre sí; y *separación* - los genes de grupos diferentes tienen una expresión de baja similitud entre sí.

Los ejemplos de diversas técnicas de agrupamiento que se han usado para el análisis de la expresión génica incluyen agrupamiento jerárquico (Eisen et al., 1998, *supra*; Alizadeh et al., 2000, *supra*; Perou et al., 2000, *supra*; Ross et al., 2000, *supra*), agrupamiento mediante el algoritmo de K medias [Herwig et al., 1999, *supra*; Tavazoie et al., 1999, Nature Genetics 22 (3), páginas 281-285], afeitado génico (Hastie et al., 2000, Genome Biology, 1 (2), Research 0003.1-0003.21), agrupamiento de bloques (Tibshirani et al., 1999, Tech. Report Univ. Stanford), el modelo de Plaid (Lazzeroni, 2002, Stat. Sinica 12, páginas 61-86) y mapas autoorganizativos (Tamayo et al., 1999, *supra*). Además, los métodos relacionados de análisis estadístico multivariante, tales como aquellos en que se utiliza la descomposición en valores singulares [Alter et al., 2000, PNAS, 97 (18), páginas 10.101-10.106; Ross et al., 2000, *supra*] o el escalamiento multidimensional, pueden ser eficaces para reducir las dimensiones de los objetos bajo estudio.

Sin embargo, métodos tales como el análisis de grupos y la descomposición en valores singulares son puramente exploratorios y sólo proporcionan una amplia visión general de la estructura interna presente en los datos. Son planteamientos no supervisados en que, en el análisis, no se utiliza la información disponible relativa a la naturaleza de la clase bajo investigación. A menudo se conoce la naturaleza de la perturbación biológica a la que se ha sometido una muestra concreta. Por ejemplo, se sabe a veces si la muestra cuyo patrón de expresión génica está siendo analizado procede de un individuo enfermo o sano. En dichos casos, se puede utilizar un análisis discriminante para clasificar las muestras en diversos grupos basándose en sus datos de expresión génica.

En dicho análisis, se construye el clasificador entrenando los datos que permiten discriminar entre miembros y no miembros de una clase dada. Luego se puede usar el clasificador entrenado para predecir la clase de muestras desconocidas. Los ejemplos de métodos de discriminación que se han descrito en la bibliografía incluyen Máquinas de Soporte Vectorial (Brown et al., 2000, PNAS 97, páginas 262-267), Vecino Más Próximo (Dudoit et al., 2000, *supra*), Árboles de Clasificación (Dudoit et al., 2000, *supra*), Clasificación Votada (Dudoit et al., 2000, *supra*), Votación Génica Ponderada (Golub et al., 1999, *supra*) y Clasificación Bayesiana (Keller et al., 2000, Tech. Report Univ. Washington). Además, se ha descrito recientemente (Nguyen y Rocke, 2002, Bioinformatics 18, páginas 39-50 y 1216-1226) una técnica en que se utiliza primero un análisis de regresión por mínimos cuadrados parciales (PLS) para reducir las dimensiones del conjunto de datos de expresión génica, seguido de una clasificación utilizando un análisis discriminante logístico (LD; del inglés, logistic discriminant) y un análisis discriminante cuadrático (QDA; del inglés, quadratic discriminant analysis).

Un desafío que plantean los datos de expresión génica a los métodos discriminatorios clásicos es que el número de genes cuya expresión está siendo analizada es muy grande en comparación con el número de muestras que se analizan. Sin embargo, en la mayoría de los casos, sólo una pequeña fracción de estos genes es informativa en los problemas del análisis discriminante. Además, existe el peligro de que el ruido de genes irrelevantes pueda enmascarar o distorsionar la información de los genes informativos. Se han sugerido varios métodos en la bibliografía para identificar y seleccionar genes que sean informativos en estudios con microrredes, tales como, por ejemplo, estadísticas t (Dudoit et al., 2002, J. Am. Stat. Ass. 97, páginas 77-87), análisis de la varianza (Kerr et al., 2000, PNAS 98, páginas 8961-8965), análisis de vecinos (Golub et al., 1999, *supra*), relación entre grupos y suma de cuadrados en los grupos (Dudoit et al., 2002, *supra*), calificación no paramétrica (Park et al., 2002, Pacific Symposium on Biocomputing, páginas 52-63) y selección de probabilidades (Keller et al., 2000, *supra*).

En los métodos aquí descritos, los datos de expresión génica que han sido normalizados y estandarizados son analizados utilizando la regresión por mínimos cuadrados parciales (PLSR). Aunque la PLSR es esencialmente un método utilizado para el análisis de datos continuos por regresión (véase el Apéndice A), puede ser también utilizado como un método para construcción de modelos y análisis discriminante utilizando una matriz de respuestas ficticia basada en una codificación binaria. La asignación de clases se basa en una simple distinción dicotómica tal como cáncer de mama (clase 1)/individuo sano (clase 2), o una distinción múltiple basada en múltiples diagnósticos de enfermedad tales como cáncer de mama (clase 1)/cáncer ovárico (clase 2)/individuo sano (clase 3). Se puede aumentar la lista de enfermedades para clasificación dependiendo de las muestras disponibles que correspondan a otros cánceres o fases de los mismos.

A la PLSR aplicada como un método de clasificación se hace referencia como PLS-DA (en que DA significa análisis discriminante; del inglés, discriminant analysis). El PLS-DA es una extensión del algoritmo PLSR en que la matriz Y es una matriz ficticia que contiene *n* filas (que corresponden al número de muestras) y *K* columnas (que corresponden al número de clases). La matriz Y se construye insertando 1 en la columna kº y -1 en todas las demás columnas si el correspondiente iº objeto de X pertenece a la clase *k*. Por regresión de Y sobre X, se consigue la clasificación de una nueva muestra seleccionando el grupo que corresponde al componente más grande del ajustado, $\hat{y}(x) = [\hat{y}_1(x), \hat{y}_2(x),..., \hat{y}_k(x)]$. De esta manera, en una matriz de respuestas -1/1, un valor de predicción inferior a 0 significa que la muestra pertenece a la clase designada como -1, mientras que un valor de predicción superior a 0 implica que la muestra pertenece a la clase designada como 1.

Una ventaja del PLSR-DA es que los resultados obtenidos pueden ser fácilmente representados en forma de dos gráficos diferentes, los gráficos de calificación y de carga. Los gráficos de calificación representan una proyección de las muestras sobre los componentes principales y muestran la distribución de las muestras en el modelo de clasificación y su relación entre sí. Los gráficos de carga presentan correlaciones entre las variables presentes en el conjunto de datos.

Normalmente se recomienda utilizar el PLS-DA como punto de partida para el problema de la clasificación a causa de su capacidad para manejar datos colineales y de la propiedad de la PLSR como técnica para reducción de dimensiones. Una vez satisfecho este objetivo, es posible utilizar otros métodos, tal como el análisis discriminante lineal (LDA; del inglés, linear discriminant analysis), que ha demostrado ser eficaz para extraer información adicional (Indahl et al., 1999, Chem. and Intell. Lab. Syst. 49, páginas 19-31). Este planteamiento se basa en descomponer primero los datos usando PLS-DA, y usar luego los vectores de calificaciones (en vez de las variables originales) como entrada para el LDA. En Duda y Hart (Classification and Scene Analysis, 1973, Wiley, EE.UU.) se pueden encontrar detalles adicionales sobre el LDA.

La siguiente operación después de la construcción del modelo es la validación del modelo. Se considera que esta operación está entre los aspectos más importantes del análisis multivariante y examina la "bondad" del modelo de calibración que se ha construido. En este trabajo, se ha usado para la validación un planteamiento de validación cruzada. En este planteamiento, se dejan una o unas pocas muestras fuera de cada segmento mientras el modelo es construido usando una validación cruzada completa sobre la base de los datos restantes. Las muestras dejadas fuera son luego utilizadas para predicción/clasificación. La repetición del proceso de validación cruzada sencilla varias veces dejando fuera diferentes muestras en cada validación cruzada conduce al llamado procedimiento de validación cruzada doble. Se ha mostrado que este planteamiento funciona bien con una cantidad limitada de datos, como es el caso de algunos de los Ejemplos aquí descritos. Además, puesto que la operación de validación cruzada se repite varias veces, se reducen los peligros de sobreajuste y sesgo del modelo.

Una vez que se ha construido y validado un modelo de calibración, los genes que presentan un patrón de expresión que es más relevante para describir la información deseada en el modelo pueden ser seleccionados mediante técnicas descritas en el campo técnico previo para selección de variables, como se menciona en otra parte. La selección de variables ayudará a reducir la complejidad final del modelo, proporcionar un modelo reducido y, por lo tanto, conducir a un modelo fiable que pueda ser usado para predicción. Además, el uso de menos genes para el fin de proporcionar un diagnóstico reducirá el costo del producto diagnóstico. De esta manera, se pueden identificar sondas informativas que se unirían a los genes de relevancia.

Hemos hallado que, una vez que se ha construido un modelo de calibración, se pueden usar eficazmente técnicas estadísticas tales como Jackknife (Effron, 1982, "The Jackknife, the Bootstrap and other resampling plans", Society for Industrial and Applied Mathematics, Philadelphia, EE.UU.), basada en una metodología de remuestreo, para seleccionar o confirmar variables significativas (sondas informativas). La varianza de incertidumbre aproximada de los coeficientes B de la regresión por PLS puede ser estimada mediante:

$$S^2B = \sum_{m=1}^{M}((B - B_m)g)^2$$

en que S^2B = varianza de incertidumbre estimada de B;

B = el coeficiente de regresión en el rango A cruzadamente validado, usando todos los N objetos;

B_m = el coeficiente de regresión en el rango A usando todos los objetos salvo el(los) objeto(s) dejado(s) fuera en el segmento m de validación cruzada; y

g = coeficiente de escalamiento (aquí, g = 1).

En nuestro planteamiento, se ha implementado Jackknife junto con la validación cruzada. Para cada variable, se calcula primero la diferencia entre los coeficientes-B B_i en un submodelo cruzadamente validado y B_{tot} para el modelo total. Luego se calcula la suma de los cuadrados de las diferencias en todos los submodelos para obtener una expresión de la varianza de la estimación de B_i para una variable. La significación de la estimación de B_i se calcula utilizando la prueba t. De este modo, los coeficientes de regresión resultantes pueden ser presentados con límites de incertidumbre que corresponden a 2 desviaciones estándares, y a partir de ello se detectan variables significativas.

No se proporcionan aquí más detalles en cuanto a la implementación o el uso de esta operación ya que esto ha sido implementado en un software comercialmente asequible, The Unscrambler, CAMO ASA, Noruega. Además, en Westad y Martens (2000, J. Near Inf. Spectr. 8, páginas 117-124) se pueden hallar detalles sobre la selección de variables utilizando Jackknife.

Se puede utilizar el planteamiento siguiente para seleccionar sondas informativas a partir de un conjunto de datos de expresión génica:

a) dejar fuera una muestra única (incluyendo sus repeticiones, si están presentes en el conjunto de datos) por segmento de validación cruzada;

b) construir un modelo de calibración (segmento cruzadamente validado) sobre las muestras restantes utilizando PLSR-DA;

c) seleccionar los genes significativos para el modelo de la operación b) utilizando el criterio Jackknife;

d) repetir las 3 operaciones anteriores hasta que todas las muestras únicas del conjunto de datos queden fuera una vez [como se describe en la operación a)]. Por ejemplo, si están presentes 75 muestras únicas en el conjunto de datos, se construyen 75 modelos de calibración diferentes para dar lugar a una colección de 75 conjuntos diferentes de sondas significativas; y

e) seleccionar las variables más significativas utilizando el criterio de frecuencia de aparición en los conjuntos generados de sondas significativas de la operación d). Por ejemplo, un conjunto de sondas que aparecen en todos los conjuntos (100%) son más informativas que las sondas que aparecen en sólo el 50% de los conjuntos generados de la operación d).

Una vez que se han seleccionado las sondas informativas para una enfermedad, se crea y valida un modelo final. Los dos modos más comúnmente usados para validar el modelo son la validación cruzada (CV; del inglés, cross-validation) y la validación del conjunto de ensayo. En la validación cruzada, los datos se dividen en k subconjuntos. El modelo es luego entrenado k veces, dejando cada vez uno de los subconjuntos fuera del entrenamiento pero utilizando sólo el subconjunto omitido para computar el criterio de error, el error cuadrático medio de predicción (RMSEP; del inglés, root mean square error of prediction). Si k es igual al tamaño de la muestra, a esto se llama validación cruzada "dejando uno fuera". La idea de dejar fuera una muestra o unas pocas muestras por segmento de validación es válida sólo en los casos en que la covarianza entre los diversos experimentos es cero. De esta manera, el planteamiento de una muestra cada vez no se puede justificar en situaciones que contienen duplicados ya que dejar fuera sólo uno de los duplicados introducirá un sesgo sistemático en nuestro análisis. En este caso, el planteamiento correcto será dejar fuera todos los duplicados de las mismas muestras a la vez ya que esto satisfaría las suposiciones de covarianza cero entre los segmentos de CV.

El segundo planteamiento para la validación del modelo es usar un conjunto de ensayo separado para validar el modelo de calibración. Esto requiere desarrollar un conjunto separado de experimentos que va a ser utilizado como un conjunto de ensayo. Éste es el planteamiento preferido dado que se dispone de datos de ensayo reales.

Luego se utiliza el modelo final para identificar el cáncer de mama en muestras de ensayo. Para este fin, se generan datos de expresión de genes informativos seleccionados a partir de las muestras de ensayo y luego se utiliza el modelo final para determinar si una muestra pertenece a una clase enferma o no enferma o tiene cáncer de mama.

Se genera preferiblemente un modelo con fines de clasificación utilizando los datos relativos a las sondas identificadas de acuerdo con el método anteriormente descrito. Preferiblemente, la muestra es como se describió previamente. Preferiblemente, los oligonucleótidos que se inmovilizan en la operación (a) son aleatoriamente seleccionados de la familia anteriormente descrita pero, alternativamente, pueden ser seleccionados para que representen las diferentes familias, por ejemplo, seleccionando uno o más de los oligonucleótidos que corresponden a genes que codifican proteínas con funciones comunes en las diferentes familias, pero que incluyen al menos las sondas del conjunto del invento. Especialmente, dicha selección se hace preferiblemente para abarcar oligonucleótidos derivados de genes de las familias (i) y (ii). Dichos oligonucleótidos pueden tener una longitud considerable, por ejemplo, si se utiliza cDNA (que está incluido dentro del alcance del término "oligonucleótido"). La identificación de dichas moléculas de cDNA como sondas útiles

permite el desarrollo de oligonucleótidos más cortos que reflejan la especificidad de las moléculas de cDNA pero son más fáciles de fabricar y manipular.

El modelo anteriormente descrito puede ser luego usado para generar y analizar datos de muestras de ensayo y, por lo tanto, se puede utilizar para los métodos diagnósticos del invento. En dichos métodos, los datos generados a partir de la muestra de ensayo proporcionan el conjunto de datos de expresión génica, y éste es normalizado y estandarizado como se describió anteriormente. Luego se ajusta éste al modelo de calibración anteriormente descrito para obtener la clasificación.

Como se mencionó previamente, a la vista del elevado contenido de información de la mayoría de los transcritos, se puede simplificar drásticamente la identificación y selección de sondas informativas para uso en el diagnóstico, control o identificación de un cáncer particular o una fase del mismo. De este modo, se puede reducir radicalmente la colección de genes a partir de la cual se puede hacer una selección para identificar sondas informativas.

A diferencia de las tecnologías de la técnica previa en que se seleccionan sondas informativas a partir de una población de miles de genes que se están expresando en una célula, como en una microrred, en el método aquí descrito se seleccionan las sondas informativas a partir de un número limitado de genes como los descritos en las familias de secuencias génicas anteriormente descritas pero que incluyen las sondas enumeradas en conjuntos del invento.

Para identificar genes que se expresan en una cantidad elevada o moderada para uso en métodos del invento, la información acerca del nivel relativo de sus transcritos en muestras de interés puede ser generada utilizando varias técnicas del campo técnico previo. Para este fin, se pueden utilizar tanto métodos no basados en secuencias, tales como la presentación diferencial y la huella dactilar de RNA, como métodos basados en secuencias, tales como microrredes o macrorredes. Alternativamente, se pueden diseñar secuencias cebadoras específicas para genes elevada y moderadamente expresados y se pueden usar métodos tales como la RT-PCR cuantitativa para determinar los niveles de los genes elevada y moderadamente expresados. Por consiguiente, un facultativo experto puede utilizar una diversidad de técnicas que son conocidas en el campo técnico para determinar el nivel relativo de mRNA en una muestra biológica.

Especial y preferiblemente, la muestra para el aislamiento de mRNA en el método anteriormente descrito es como se describió previamente y, preferiblemente, no es del sitio de la enfermedad, y las células de dicha muestra no son células de la enfermedad y no han entrado en contacto con células de la enfermedad, tal como, por ejemplo, el uso de una muestra de sangre periférica para la detección de cáncer de mama.

Los ejemplos siguientes se proporcionan sólo a modo de ilustración, en los que las Figuras a que se hace referencia son las siguientes:

La Figura 1 muestra la posible interacción de varios factores responsables de cambios en la expresión en un individuo con cáncer de mama;

La Figura 2 muestra la proyección de 102 muestras normales (incluyendo benignas) y de cáncer de mama sobre un modelo de clasificación generado mediante PLSR-DA usando los datos de 35 genes informativos, en que PC (del inglés, principal components) son los componentes principales y N y C son muestras normales y de cáncer de mama, respectivamente.

La Figura 3 muestra un gráfico de predicción basado en tres componentes principales utilizando los datos de 35 cDNAs; y

La Figura 4 muestra el nivel medio de expresión de los 35 genes utilizados para la predicción de cáncer de mama.

Ejemplo 1: Diagnóstico del cáncer de mama

Métodos

Muestras sanguíneas

Se recogieron muestras sanguíneas de donantes con su consentimiento informado bajo una aprobación del Comité Ético Regional de Noruega. Todos los donantes fueron anónimamente tratados durante el análisis. Se extrajo sangre de hembras con un mamograma inicial sospechoso, que incluía tanto hembras con cáncer de mama como hembras con mamogramas anormales, antes de cualquier conocimiento sobre si la anormalidad observada durante la primera exploración era benigna o maligna. En todos los casos, las muestras sanguíneas se extrajeron entre las 8 de la mañana y las 4 de la tarde. Personal experto extrajo de cada hembra 10 ml de sangre en tubos Vacutainer que contenían EDTA como anticoagulante (Becton Dickinson, Baltimore, EE.UU.) o directamente en tubos PAXgene™ (PreAnalytiX, Hombrechtikon, Suiza). La sangre recogida en los tubos con EDTA se almacenó inmediatamente a -80 °C, mientras que los tubos PAX se dejaron durante la noche y después se almacenaron a -80 °C hasta ser utilizados.

Preparación de redes de cDNA

Se escogieron aleatoriamente 1435 clones de cDNA de un banco plasmídico construido a partir de sangre completa de

550 individuos sanos (Clontech, Palo Alto, EE.UU.). Aproximadamente el 20% de los clones aleatoriamente escogidos eran redundantes. Para la multiplicación de los insertos, se cultivaron clones bacterianos en placas para microtitulación que contenían 150 µl de LB con 50 µg/ml de carbenicilina y se incubaron durante la noche a 37 °C con agitación. Para lisar las células, se diluyeron 5 µl de cada cultivo con 50 µl de H_2O y se incubó la mezcla a 95 °C durante 12 minutos. De esta mezcla, se sometieron 2 µl a una reacción PCR usando 40 micromoles de cebador de secuenciación 5' y 3' en presencia de $MgCl_2$ 1,5 mM. Las reacciones PCR se llevaron a cabo con el siguiente protocolo de ciclos: 4 minutos a 95 °C, seguido de 25 ciclos de 1 minuto a 94 °C, 1 minuto a 60 °C y 3 minutos a 72 °C en un termociclador RoboCycler® (Stratagene, La Jolla, EE.UU.) o un termociclador Peltier "DNA Engine Dyad" (MJ Research Inc., Waltham, EE.UU.). Los productos multiplicados fueron desnaturalizados con NaOH (concentración final de 0,2 M) durante 30 minutos y salpicados sobre membranas Hybond-N^+ (Amershem Pharmacia Biotech, Little Chalfont, Reino Unido), utilizando una terminal de trabajo MicroGrid II de acuerdo con las instrucciones del fabricante (BioRobotics Ltd., Cambridge, Inglaterra). Los cDNAs inmovilizados se fijaron utilizando un agente entrecruzante por luz UV (Hoefer Scientific Instruments, San Francisco, EE.UU.).

Además de los 1435 cDNAs, las redes impresas también contenían testigos para evaluar el nivel de fondo, la consistencia y la sensibilidad del ensayo. Estos fueron salpicados en múltiples posiciones e incluían testigos tales como mezcla para PCR (sin inserto alguno); testigos del sistema de validación de red SpotReport™ 10 (Stratagene, La Jolla, EE.UU.) y cDNAs correspondientes a genes constitutivamente expresados tales como los de β-actina, γ-actina, GAPDH, HOD y ciclofilina.

Extracción del RNA, síntesis de sondas e hibridación

La sangre recogida en los tubos con EDTA fue descongelada a 37 °C y transferida a tubos PAX, y el RNA total fue purificado de acuerdo con las instrucciones del proveedor (PreAnalytiX, Hombrechtikon, Suiza). De la sangre recogida directamente en tubos PAX, se extrajo el RNA total en los tubos del modo anterior sin transferencia alguna a nuevos tubos. El DNA contaminante fue eliminado del RNA aislado mediante un tratamiento con DNasa I usando un kit exento de DNA (Ambion, Inc., Austin, EE.UU.). Se determinó visualmente la calidad del RNA inspeccionando la integridad de las bandas ribosómicas 28S y 18S después de una electroforesis en gel de agarosa. En este estudio sólo se utilizaron las muestras de las que se extrajo RNA de buena calidad. Según nuestra experiencia, la sangre recogida en tubos con EDTA daba lugar a RNA de mala calidad, mientras que la recogida en tubos PAX casi siempre proporcionaba RNA de buena calidad. Se determinaron la concentración y la pureza del RNA extraído, midiendo la absorbancia a 260 nm y 280 nm. Del RNA total se aisló mRNA utilizando Dynabeads de acuerdo con las instrucciones del proveedor (Dynal AS, Oslo, Noruega).

Se llevaron a cabo experimentos de etiquetado e hibridación en 16 lotes. El número de muestras examinadas en cada lote variaba de seis a nueve. Para minimizar el ruido debido a la variación de lote a lote en la impresión, en cada lote sólo se usaron las redes fabricadas durante el mismo proceso de impresión. Cuando se examinaron muestras más de una vez (duplicados), se utilizaron partes alícuotas de la misma colección de mRNA para la síntesis de las sondas. Para la síntesis de las sondas, partes alícuotas de mRNA correspondientes a 4-5 µg de RNA total fueron mezcladas con oligodT_{25NV} (0,5 µg/µl) y puntas ("spikes") de mRNA del sistema de validación de red SpotReport™ 10 (10 pg; punta 2, 1 pg), calentadas a 70 °C y luego enfriadas sobre hielo. Se prepararon las sondas en mezclas de reacción de 35 µl mediante transcripción inversa en presencia de 1850 kBq de [α^{33}P]-dATP, dATP 3,5 µM, 0,6 mM de cada uno de dCTP, dTTP y dGTP, 200 unidades de transcriptasa inversa SuperScript (Invitrogen, Life Technologies) y DTT 0,1 M, realizándose el etiquetado durante 1,5 horas a 42 °C. Después de la síntesis, se desactivó la enzima durante 10 minutos a 70 °C y se separó el mRNA incubando la mezcla de reacción durante 20 minutos a 37 °C en 4 unidades de Ribo H (Promega, Madison, EE.UU.). Los nucleótidos no incorporados fueron separados usando columnas ProbeQuant G 50 (Amersham Biosciences, Piscataway, EE.UU.).

Se equilibraron las membranas en SSC 4x durante 2 horas a temperatura ambiental y se prehibridaron durante la noche a 65 °C en 10 ml de disolución de prehibridación (SSC 4x, NaH_2PO_4 0,1 M, EDTA 1 mM, sulfato de dextrano al 8%, disolución de Denhardt 10x, SDS al 1%). Se añadieron sondas recién preparadas a 5 ml de la misma disolución de prehibridación y se continuó la hibridación durante la noche a 65 °C. Se lavaron las membranas a 65 °C con rigor creciente (2 x 30 minutos, cada vez en SSC 2x, SDS al 0,1%; SSC 1x, SDS al 0,1%; SSC 0,1x, SDS al 0,1%).

Cuantificación de las señales de hibridación

Se expusieron las membranas hibridadas a Phosphoscreen (superresolución) durante dos días y se generó un archivo de imágenes usando PhosphoImager (Cyclone, Packard, Meriden, EE.UU.). La identificación y cuantificación de las señales de hibridación, así como la resta de los valores locales de fondo, se llevaron a cabo utilizando el software Phoretix (Non Linear Dynamics, Reino Unido). Para la resta del fondo, la mediana de la línea de píxeles alrededor del contorno de cada mancha fue restada de la intensidad de las señales evaluadas en cada mancha.

Análisis de datos

De los 1435 datos de expresión con el fondo restado, se eliminaron las señales de 67 genes de cada membrana para excluir los genes expresados con un alto grado de varianza. Esto incluía la eliminación del 1,25% de las señales más bajas y más altas de cada membrana. Para la normalización, se dividió primero el valor de cada mancha por la media de

las señales de cada red, lo que fue seguido de una transformación por raíz cúbica de todas las manchas. Los datos normalizados fueron luego ajustados por lotes usando un análisis de la varianza de una vía (ANOVA).

Los datos preprocesados fueron luego utilizados para aislar las sondas informativas por:

a) construcción de un modelo de PLSR cruzadamente validado, en el que se dejó fuera una muestra única (incluyendo todas las repeticiones de la muestra seleccionada) por segmento de validación cruzada;

b) selección del conjunto de genes significativos para el modelo de la operación a) utilizando el criterio Jackknife;

c) construcción de un modelo de PLSR-DA cruzadamente validado como en la operación a) utilizando el gen seleccionado en la operación b);

d) selección de nuevo del conjunto de genes más significativos para el modelo de la operación c) utilizando el criterio Jackknife.

La operación b) dio lugar a 125 genes.

La operación d) dio lugar a la selección de 35 genes significativos. Se construyó un modelo de clasificación final basándose en estos genes.

Las sondas informativas seleccionadas basadas en el criterio de aparición fueron utilizadas para construir un modelo de clasificación. En la Figura 2 se muestra el resultado del modelo de clasificación basado en 35 sondas, en el que se ve que el patrón de expresión de estos genes permitía clasificar a la mayoría de las mujeres con cáncer de mama y las mujeres sin cáncer de mama en grupos distintos. En esta figura, PC1 y PC2 indican los dos componentes principales estadísticamente derivados de los datos que mejor definen la variabilidad sistémica presente en los datos. Esto permite que cada muestra, y los datos de cada una de las sondas informativas a las que se unió el cDNA de primera cadena etiquetado de la muestra, se representen sobre el modelo de clasificación como un solo punto que es una proyección de la muestra sobre los componentes principales - el gráfico de calificación.

La Figura 3 muestra el gráfico de predicción utilizando los 35 genes significativos. En el gráfico de predicción mostrado, las muestras con cáncer aparecen sobre el eje X en +1 y las muestras sin cáncer aparecen en -1. El eje Y representa el predicho miembro de clase. Durante la predicción, si la predicción es correcta, las muestras con cáncer deberían caer por encima de cero y las muestras sin cáncer deberían caer por debajo de cero. En cada caso, casi todas las muestras son correctamente predichas. Para la validación cruzada, se dividieron 102 muestras experimentales en 60 segmentos de validación cruzada, en que cada segmento representaba una muestra única e incluía sus duplicados, si estaban presentes.

Se consiguió la predicción correcta de la mayoría de las células con cáncer de mama. Se predijeron correctamente 19 de 22 pacientes con cáncer, así como 34/35 pacientes normales. En la Tabla 1 se muestran los detalles completos de los individuos examinados y la exactitud de la predicción. En la Tabla 2 se proporcionan detalles de los 35 genes informativos, los genes de bases de datos públicas con los que muestran similitud secuencial, y su supuesta función biológica. Sus secuencias vienen después de estos ejemplos.

En la Figura 4 se muestra el nivel de expresión de los 35 genes, y se verá que algunos están sobreexpresados y otros infraexpresados con respecto a la expresión en pacientes normales.

Ejemplo 2: Identificación de otras sondas informativas y uso en el diagnóstico del cáncer de mama

Métodos

Los métodos para identificación y análisis utilizados fueron esencialmente como los descritos en el Ejemplo 1 salvo por que, en lugar de preparar una red de cDNA, se analizaron las muestras utilizando una plataforma comercialmente asequible para el análisis de expresión génica a gran escala (chip Agilent 22K).

Se analizó un gran número de muestras que comprendía un total de 122 (78 testigo y 44 con cáncer de mama). Se analizaron los datos utilizando PLSR del modo previamente descrito. Los genes de interés se seleccionaron mediante un planteamiento de validación cruzada de 10 iteraciones. De este modo, los datos de las 122 muestras fueron divididos en 10 conjuntos, conteniendo cada conjunto 12-13 muestras. Se construyó un modelo de calibración sobre nueve conjuntos dejando fuera un conjunto. Los genes significativos fueron identificados mediante la técnica Jackknife sobre el modelo incorporado. Se repitieron estas operaciones para los 10 conjuntos dejando fuera cada conjunto al menos una vez. Luego se identificaron los genes informativos basándose en el criterio de frecuencia de aparición. Se halló que 109 genes eran informativos en los 10 modelos de calibración.

Resultados

Se utilizaron los 109 genes anteriormente descritos y 3 genes más para predecir la clasificación de 122 de las muestras usadas. Los resultados se muestran en la tabla siguiente.

MUESTRA	Número	Correctamente predicho	Incorrectamente predicho	Índice de error
Testigo	78	67	11	0,14
Cáncer de mama	44	26	18	0,41

Los 109 genes informativos pueden ser divididos en tres categorías; a saber, aquellos que caen en las familias (i) y (ii) que aquí se describen, y otros genes. En la Tabla 3 se proporcionan detalles de las sondas informativas cuyos genes correspondientes caen en las familias (i) y (ii) y proporciona el número asignado por Agilent a esa sonda. En la Tabla 4 se proporcionan similarmente detalles de las sondas informativas cuyos genes correspondientes no parecen caer dentro de la familia (i) ni (ii). En las Tablas 5 y 6 se proporcionan detalles de los genes con los que las sondas de las Tablas 3 y 4, respectivamente, muestran similitud de secuencias, su supuesta función biológica, cuando se conoce, y los números de acceso para esos genes.

Apéndice A

Regresión por Mínimos Cuadrados Parciales (PLSR)

Sea definido un modelo de regresión multivariante como:

$Y = XB + F$

en que

X es una matriz *NxP* con *N* variables de predicción (genes);

siendo Y (*NxJ*) las *J* variables predichas. En nuestro caso, Y representa una matriz que contiene variables ficticias;

B es una matriz de coeficientes de regresión; y

F es una matriz *NxJ* de residuales.

La estructura del modelo de PLSR puede ser descrita como:

$X = TP^T + E_A$, e

$Y = TQ^T + F_A$,

en que

T (*NxA*) es una matriz de vectores de calificación que son combinaciones lineales de las variables x;

P (*PxA*) es una matriz con los vectores p_a de carga de x como columnas;

Q (*JxA*) es una matriz con los vectores q_a de carga de y como columnas;

E_a (*NxP*) es la matriz para X después de A factores; y

F_a (*NxJ*) es la matriz para Y después de A factores.

En PLSR, el criterio es maximizar la expuesta covarianza de [X,Y]. Esto se alcanza mediante el vector w_{a+1} de pesos de carga, que es el primer vector propio de $E_a^T F_a F_a^T E_a$ (E_a y F_a son los X e Y desinflados después de *a* factores o componentes de PLS).

Se obtienen los coeficientes de regresión mediante:

$B = W(P^T W)^{-1} Q^T$

Un modelo de PLSR con rango completo, es decir, número máximo de componentes, es equivalente a las soluciones de la MLR. En Marteus y Naes, 1989, "Multivariate Calibration", John Wiley and Sons, Inc., EE.UU., y Kowalski y Seasholtz, 1991, *supra*, se pueden hallar más detalles sobre PLSR.

Secuencias nucleotídicas de 34/35 genes seleccionados mediante Jackknife

Identificaciones de clones y sus secuencias

I-30 (Secuencia 1)

CTTTTCCTCCCGCTGTCCCCCACGGAGGGGACTGCTCTCCCCCGCTGCATCCTT
TCTGTGAGGTACCTTACCCACCTCAGCACCTGAGAGGGTGAAATAGAATTCTAA
CCTCGACATTCGGGAAGTGTTTTTGAGAAGTCTCGGTCGGTAAGGGAAGTCTTC
CAAGTCCGTGCAGCACTAACGTATTGGCACCTGCCTCCTCTTCGGCCACCCCCC
AGATGAGGCAGCTGTGACTGTGTCAAGGGAAGCCACGACTCTGACCATAGTCTT
CTCTCAGCTTCCACTGCCGTCTCCACAGGAAACCCAGAAGTTCTGTGAACAAGT
CCATGCTGCCATCAAGGCATTTATTGCAGTGTACTATTTGCTTCCAAAGGATCA
GGCCCTGAGAACAATGACCTTATTTCCTACAACAGTGTCTGGGTTGCGTGCCAG
CAGATGCCTCAGATACCAAGAGATAACAAAGCTGCAGCTCTTTGATGCTGACC
AAGAATGTGGATTTTGTGAAGGATGCACATGAAGAAATGGAGCAGGCTGTGGAA
GAATGTGACCCTTACTCTGGCCTCTTGAATGATACTGAGGAGAACAACTCTGAC
AACCACAATCATGAGGATGATGTGTTGGGGTTTCCCAGCAATCAGGACTTGTAT
TGGTCAGAGGACGATCAAGAGCTCATAATCCCATGCCTTGCGCTGGTGAGAGCA
TCCAAAGCCTGCCTGAAGAAAATTCGGATGTTAGTGGCAGAGAATGGGAAGAAG
GATCAGGTGGCACAGCTGGATGACATTGTGGATATTTCTGATGAAATCAGCCCT
AGTGTGGATGATTTGGCTCTGAGCATATATCCACCTATGTGTCACCTGACCGTG
CGAATCAATTCTGCGAAACTTGTATCTGTTTTAAAGAAGGCACTTGAAATTACA
AAAGCAAGTCATGTGACCCCTCAGCCAGAAGATAGTTGGATCCCTTTACTTATT
AATGCCATTGATCATTGCATGAATAGAATCAAGGAGCTCACTCAGAGTGAACTT
GAATTATGACTTTTCAGGCTCATTTGTACTCTCTTCCCCTCTCATCGTCATGGT
CAGGCTCTGATACCTGCTTTTAAAATGGAGCTAGAATGCTTGCTGGATTGAAAG
GGAGTGCCTATCTATATTTAGCAAGAGACACTATTACCAAAGATTGTTGGTTAG
GCCAGATTGACACCTATTTATAAACCATATGCGTATATTTTTCTGTGCTATATA
TGAAAAATAATTGCATGATTTCTCATTCCTGAGTCATTTCTCAGAGATTCCTAG
GAAAGCTGCCTTATTCTCTTTTTGCAGTAAAGTATGTTGTTTTCATTGTAAAGA
TGTTGATGGTCTCAATAAAATGCTAACTTGCCAGTGAAAAAAAAAAAAAA

III-02 (Secuencia 2)

AGGATCTAAGACCAGCCTGGCAGCCACCAGATGGTGATTCTAGTCCTGGCTCAG
TCAGTAATAGGTCACTGACCCCAGAGAAATCAATTCAGCCTCCCCAGGTCCTTG
GATTTCTTTCTGTGAAAATGAAAGCATAGGTAGGAATTTCCCATGGAACAGCTA
GCAGAGGAGAAATATTAAAAGTCAGGAGACTCATGCTATAGTTTTCATACTTCA

TTACAACAATGTTGTTTAGGACAAGTGAGTTAACCTGTTAGCTTCCTCTATATA
AAATGGAAAGTCATTAAAAACCTACATAGCAGGGTTCTTGTGAAGATCAAGTGA
TAATGTAGGAAGCATGTACAAATGTCACATTCTGCCGTCACGTAATGGTCCTCA
CAGCTTGAGGTAGCATTTAGCATGTGTCATGATTTAGTACAAGGGTTGGCAAAC
TGTTGCTCTTGGATTAAGTCTGGCTCATTGCCTGTTTTTCAAAGAAAAAAATTG
TATATGTGTGTATATATGTTATATATAGGTACACACACATATGTGCTATATATA
GCATATATACACACATAATATATAAACATGTACATATATAGCATTATATATATA
CGTGTATAATATCTCCAGTCCTCATGACCAGCCATGCTTGTTCATTTACATTTG
CATACTCTATGATTGCTTTCATGCAACAATGGCAGAGTTGAGTGATTGTTTTGC
AACAGAGACTGTATGGCCCACTAAACCTAAAATATTTAGTCTCTGACCCTGAAA
TGTAAGATTGATAGCCCAGGACCAGGCGTGGTGGCTCACACTTGTAATCCTAGC
ACTTTGGCAGGCCAAGGAGGGTGGATCACCTGAGGTCAGGAGTTAAAGACCAGC
CTGGCCAACATGGTGAAACCCTGACTCTACTAAAAATACAGAAATTAGCTGGGC
GTGGTAATGGGTGCCTGCAATCCAAGCTACTCTGGAGGCTGAGGCAGGAGAATC
ACTTGAACCCAGGAGGCAGAAGTTACAGTGAGCTGAGATGGTGCCACTGCACTC
CAGCCTGGACGACAGAGTGAGACTCCATCTCAAAAA

III-27 (Secuencia 3)

CCATTCTCCTGCCTCAGCCTCTCAAGTAGCTGGGACTACAGGCGCCCACAACCA
CGCCCGGCTAATGTTTTTGGTATTTTCGTAGAGACGGGGTTTCACCTTGTTAG
CCAGGATGGTCTTGATCTCCTGACCTCGTGATCTGCCTGCCTCGGCCTCCCAAA
GTGTTGGGATTACAGGCACATTTTTCACAATTTTTTAACACTTAAGAATGACTT
AACTGAATCATGCCTTTAGAAGAAACTTTCTGTTTAAAAAAAAAAAAAAA

III-60 (Secuencia 4)

CTGCCGCCGCCCCCAGCTCCCCCGCCTCGGGGAGGGCACCAGGTCACTGCAGCC
AGAGGGGTCCAGAAGAGAGAGGAGGCACTGCCTCCACTACAGCAACTGCACCCA
CGATGCAGAGCATCAAGTGCGTGGTGGTGGGTGATGGGGCTGTGGGCAAGACGT
GCCTGCTCATCTGCTACACAACTAACGCTTTCCCCAAAGAGTACATCCCCACCG
TGTTCGACAATTACAGCGCGCAGAGCGCAGTTGACGGGCGCACAGTGAACCTGA
ACCTGTGGGACACTGCGGGCCAGGAGGAGTATGACCGCCTCCGTACACTCTCCT
ACCCTCAGACCAACGTTTTCGTCATCTGTTTCTCCATTGCCAGTCCGCCGTCCT
ATGAGAACGTGCGGCACAAGTGGCATCCAGAGGTGTGCCACCACTGCCCTGATG
TGCCCATCCTGCTGGTGGGCACCAAGAAGGACCTGAGAGCCCAGCCTGACACCC
TACGGCGCCTCAAGGAGCAGGGCCAGGCGCCCATCACACCGCAGCAGGGCCAGG
CACTGGCCAAGCAGATCCACGCTGTGCGCTACCTCGAATGCTCAGCCCTGCAAC
AGGATGGTGTCAAGGAAGTGTTCGCCGAGGCTGTCCGGGCTGTGCTCAACCCCA

CGCCGATCAAGCGTGGGCGGTCCTGCATCCTCTTGTGACCCTGGCACTTGGCTT
GGAGGCTGCCCCTGCCCTCCCCCCACCAGTTGTGCCTTGGTGCCTTGTCCGCCT
CAGCTGTGCCTTAAGGACTAATTCTGGCACCCCTTTCCAGGGGGTTCCCTGAAT
GCCTTTTTCTCTGAGTGCCTTTTTCTCCTTAAGGAGGCCTGCAGAGAAAGGGGC
TTTGGGCTCTGCCCCCCTCTGCTTGGGAACACTGGGTATTCTCATGAGCTCATC
CAAGCCAAGGTTGGACCCCTCCCCAAGAGGCCAACCCAGTGCCCCCTCCCATTT
TCCGTACTGACCAGTTCATCCAGCTTTCCACACAGTTGTTGCTGCCTATTGTGG
TGCCGCCTCAGGTTAGGGGCTCTCAGCCATCTCTAACCTCTGCCCTCGCTGCTC
TTGGAATTGCGCCCCCAAGATGCTCTCTCCCTTCTCCAATGAGGGAGCCACAGA
ATCCTGAGAAGGTGAATGTGCCCTAACCTGCTCCTCTGTGCCTAGGCCTTACGC
ATTTGCTGACTGACTCAGCCCCCATGCTTCTGGGGACCTTTCCTACCCCCATCA
GCATCAATAAAACCTCCTGTCTCCAGTGA

IV-26 (Secuencia 5)

CAGCCCTCCGTCACCTCTTCACCGCACCCTCGGACTGCCCCAAGGCCCCCGCCG
CCGCTCCAGCGCCGCGCAGCCACCGCCGCCGCCGCCGCCTCTCCTTAGTCGCCG
CCATGACGACCGCGTCCACCTCGCAGGTGCGCCAGAACTACCACCAGGACTCAG
AGGCCGCCATCAACCGCCAGATCAACCTGGAGCTCTACGCCTCCTACGTTTACC
TGTCCATGTCTTACTACTTTGACCGCGATGATGTGGCTTTGAAGAACTTTGCCA
AATACTTTCTTCACCAATCTCATGAGGAGAGGGAACATGCTGAGAAACTGATGA
AGCTGCAGAACCAACGAGGTGGCCGAATCTTCCTTCAGGATATCAAGAAACCAG
ACTGTGATGACTGGGAGAGCGGGCTGAATGCAATGGAGTGTGCATTACATTTGG
AAAAAAAATGTGAATCAGTCACTACTGGAACTGCACAAACTGGCCACTGACAAAA
ATGACCCCCATTTGTGTGACTTCATTGAGACACATTACCTGAATGAGCAGGTGA
AAGCCATCAAAGAATTGGGTGACCACGTGACCAACTTGCGCAAGATGGGAGCGC
CCGAATCTGGCTTGGCGGAATATCTCTTTGACAAGCACACCCTGGGAGACAGTG
ATAATGAAAGCTAAGCCTCGGGCTAATTTCCCCATAGCCGTGGGGTGACTTCCC
TGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGT
TGTACCAAAACATCCACTTAAGTTCTTTGATTTGTACCATTCCTTCAAATAAAG
AAATTTGGTACCCAAAAAAAA

IV-41 (Secuencia 6)

GCCATTTCTAAGACCTACAGCTACCTGACCCCCGACCTCTGGAAGGAGACTGTA
TTCACCAAGTCTCCCTATCAGGAGTTCACTGACCACCTCGTCAAGACCCACACC
AGAGTCTCCGTGCAGCGGACTCAGGCTCCAGCTGTGGCTACAACATAGGGTTTT
TATACAAGAAAAATAAAGTGAATTAAGCGTGAAAA

IV-51 (Secuencia 7)

ATTTCTGTGGATACAGTGCCCACCGCCCTCCTCCACTTGGAAACGGTATCCTCC
CTGCCCATCCGTCTGTCTGTCGCCCTTCTCCCGGCCCTCACTAAGCCCCGGCAC
TTCTAGTGGTCTCACCTGGAGGCAAGAGGGAGGGGACAGAGGCCCTGCCACGTC
CCGCTGCCTCCTGCTCTCTGGAGGTACTGAGACAGGGTGCTGATGGGAAGGAGG
GGAGCCTTTGGGGGGCCACCCGGGGCCTGGACCTATGCAGGGAGGCCACGTCCC
ACCCCACCTCTTGTTTCTGGGTCCCTGCTCCCCTTTGGGGGTGTGTGTGTGTGT
TTTAATTTTCTTTATGGAAAAATTGACAAAAAAAAATAGAGAGAGAGGTATTTA
ACTGCAATAAACTGGCCCCATGTGGCCCCCGCCTTGTCAAAAAAAAAA

V-09 (Secuencia 8)

TGGATTCCCGTCGTAACTTAAAGGGAAACTTTCACAATGTCCGGAGCCCTTGAT
GTCCTGCAAATGAAGGAGGAGGATGTCCTTAAGTTCCTTGCAGCAGGAACCCAC
TTAGGTGGCACCAATCTTGACTTCCAGATGGAACAGTACATCTATAAAAGGAAA
AGTGATGGCATCTATATCATAAATCTCAAGAGGACCTGGGAGAAGCTTCTGCTG
GCAGCTCGTGCAATTGTTGCCATTGAAAACCCTGCTGATGTCAGTGTTATATCC
TCCAGGAATACTGGCCAGAGGGCTGTGCTGAAGTTTGCTGCTGCCACTGGAGCC
ACTCCAATTGCTGGCCGCTTCACTCCTGGAACCTTCACTAACCAGATCCAGGCA
GCCTTCCGGGAGCCACGGCTTCTTGTGGTTACTGACCCCAGGGCTGACCACCAG
CCTCTCACGGAGGCATCTTATGTTAACCTACCTACCATTGCGCTGTGTAACACA
GATTCTCCTCTGCGCTATGTGGACATTGCCATCCCATGCAACAACAAGGGAGCT
CACTCAGTGGGTTTAATGTGGTGGATGCTGGCTCGGGAAGTTCTGCGCATGCGT
GGCACCATTTCCCGTGAACACCCATGGGAGGTCATGCCTGATCTGTACTTCTAC
AGAGATCCTGAAGAGATTGAAAAAGAAGAGCAGGCTGCTGCTGAGAAGGCAGTG
ACCAAGGAGGAATTTCAGGGTGAATGGACTGCTCCCGCTCCTGAGTTCACTGCT
ACTCAGCCTGAGGTTGCAGACTGGTCTGAAGGTGTACAGGTGCCCTCTGTGCCT
ATTCAGCAATTCCCTACTGAAGACTGGAGCGCTCAGCCTGCCACGGAAGACTGG
TCTGCAGCTCCCACTGCTCAGGCCACTGAATGGGTAGGAGCAACCACTGACTGG
TCTTAAGCTGTTCTTGCATAGGCTCTTAAGCAGCATGGAAAAATGGTTGATGGA
AAATAAACATCAGTTTCT

V-38 (Secuencia 9)

GTTTAAATTTGACAAACTAAAGCTAATTACTGCTATAAGAGTAATAACTGCTCA
TTTTCCATAACTCATTCTTAAAGTTTTAGTAATGTAAAAGTTATTTTTTTGCAG
TAAGTTATAATGATAGAAGCTTACATGTTTTTTCATGCCTCATCTGTTTCCCCT
TAAAACTATAATTATCAGTAAAGTCCTGTGGTATTTTTCAATTTGTAAGAAACT
AGGCTATATATACATTGGGAAAAACAGCCTTCATTTGTCAATGCACTAGTGTTC
CAAAGGTTTCTGGTAATTGTGTGCTATTGCTTTTTGTTGACTTGCAAAAAAAAA

AAAAAAAAAATTACTATGACTTGTGGTAGCCCTGCAACCTTCGGAAGTGCTTAG
CCCAGTCTGACCATACATTTATATTTAGAATGCTTAGGTAAATAAATAATATGC
CTAAACCCAATGCTATAAGATACTATATAATATCTCATAATTTTAAAAATCACT
GTTTTGTATAATAATAAAACAAGGCAGGCAAGCTGTTCTACAATGACTGTTGGT
AAGGGTGCTGAGGAAGAAAAACAAACAATCTTGATTCAGGGATAGTGAATAGAC
AAAAAATGTCCTAATCAATGAAGCTGTGTGATGATTCTGATTGACAGAGAGTGC
TGCCACAAGATTCTTAGGCTACACTCAAATCAGCAGAAAAAGTGCTACAATAAA
TTAGAAGTGACTATTACAGGTGCAGATGAGGGTTGGTAGTACCTGTTTGCCATT
TCTCTTCTAATCTTATATTTCTGACCCTCCTACTGTAAGTCGCGCGGAGGCGG
AGGCTTGGGTGCGTTCAAGATTCAACTTCACCCGTAACCCACCGCCATGGCCGA
GGAAGGCATTGCTGCTGGAGGTGTAATGGACGTTAATACTGCTTTACAAGAGGT
TCTGAAGACTGCCCTCATCCACGATGGCCTAGCACGTGGAATTCGCGAAGCTGC
CAAAGCCTTAGACAAGCGCCAAGCCCATCTTTGTGTGCTTGCATCCAACTGTGA
TGAGCCTATGTATGTCAAGTTGGTGGAGGCCCTTTGTGCTGAACACCAAATCAA
CCTAATTAAGGTTGATGACAACAAGAAACTAGGAGAATGGGTAGGCCTTTGTAA
AATTGACAGAGAGGGGAAACCCCGTAAAGTGGTTGGTTGCAGTTGTGTAGTAGT
TAAGGACTATGGCAAGGAGTCTCAGGCCAAGGATGTCATTGAAGAGTATTTCAA
ATGCAAGAAATGAAGAAATAAATCTTTGGCTCACAAA

VI-44 (Secuencia 10)

GAGAATGGCTTGAACCCAGTAGGCAGAGGTTGTAGTGAGCCGAGATTGGGCCAC
TGCACTTTAGCCTGGGTGACAGAGTGAGACTCTGTCTCAAAAAAAAAAAAAAAA
AATTTAAATAAAATAAAAAACCTTTACTTATTTTTAAATTGGGTTGTCTTTTTG
GTATTGAGTTGTTAAAGTTCTTTATATATTTTAGGTACAAATCCCTTATGAGAT
ACGTGATTTGAAAATATTTTCTCCCATTCTGTGGGTTGCTTTTTCACTTTCTTG
GTTGTATCCTTTGAAGCACAGAAGTTTTAAATTTTGATGAAGTCCAGTTTATTT
ATTTTTTTGCTGTTGTTTCTGCTCATACTTTTGAGGTCATGTCTGAGAAACCAT
TGTCAAATCCAAGGTCGTGATGACTTACCCCTGTGTTTTCTTCTAAGAGTTTTA
AAGGCATCTGAAGCTTAATGTGCACTAGATGGATTCTAAATATCATCTCATCCA
AAACCTGCTATATATACTACCTTCCTCATCTCAGTTGAAGGCAAGTCCATTGTT
TCAATTGCCTGGGCAAAAAATATTCTAAATAATTCATAATTTTTCCTCAACTCC
ACATCTATTGGTAAATCCTGTGGGTTCTCCTTTTAAAACATATCCAAAATAGAA
TCATTTCTCACTATCATTCCACTGCAGGCACCAAGTCTCAATAGTCTCCTAGCA
GATAATCATGTCTACATTTATTCTCAATGTAGCAGCTAGAGAGCTTTTTTG

VI-49 (Secuencia 11)

GCGGTCGTAAGGGCTGAGGATTTTTGGTCCGCACGCTCCTGCTCCTGACTCACC

GCTGTTCGCTCTCGCCGAGGAACAAGTCGGTCAGGAAGCCCGCGCGCAACAGCC
ATGGCTTTTAAGGATACCGGAAAAACACCCGTGGAGCCGGAGGTGGCAATTCAC
CGAATTCGAATCACCCTAACAAGCCGCAACGTAAAATCCTTGGAAAAGGTGTGT
GCTGACTTGATAAGAGGCGCAAAAGAAAAGAATCTCAAAGTGAAAGGACCAGTT
CGAATGCCTACCAAGACTTTGAGAATCACTACAAGAAAAACTCCTTGTGGTGAA
GGTTCTAAGACGTGGGATCGTTTCCAGATGAGAATTCACAAGCGACTCATTGAC
TTGCACAGTCCTTCTGAGATTGTTAAGCAGATTACTTCCATCAGTATTGAGCCA
GGAGTTGAGGTGGAAGTCACCATTGCAGATGCTTAAGTCAACTATTTTAATAAA
TTGATGACCAGTTGTTAAAAAAAAAAAAAA

VI-52 (Secuencia 12)

GAAAAGGGNTNGCNCCCAANGGGCAGAGGTTGGGCTGATGCCGATATTGGGCCN
CTGCNCTNCANACCTGGGTGACATGAATGAAACTCTGTCTCACATAAAAACCCA
AAAAANCTAAATGAAATAAAAGACCTTTGCTTATTNCTAANTTGGGTACGC

VII-15 (Secuencia 13)

CCCATCCCCTCGACCGCTCGCGTCGCATTTGGCCGCCTCCCTACCGCTCCAAGC
CCAGCCCTCAGCCATGGCATGCCCCCTGGATCAGGCCATTGGCCTCCTCGTGGC
CATCTTCCACAAGTACTCCGGCAGGGAGGGTGACAAGCACACCCTGAGCAAGAA
GGAGCTGAAGGAGCTGATCCAGAAGGAGCTCACCATTGGCTCGAAGCTGCAGGA
TGCTGAAATTGCAAGGCTGATGGAAGACTTGGACCGGAACAAGGACCAGGAGGT
GAACTTCCAGGAGTATGTCACCTTCCTGGGGGCCTTGGCTTTGAT

VII-32 (Secuencia 14)

AATTAGAGAGGTGAGGATCTGGTATTTCCTGGACTAAATTCCCCTTGGGGAAGA
CGAAGGGATGCTGCAGTTCCAAAAGAGAAGGACTCTTCCAGAGTCATCTACCTG
AGTCCCAAAGCTCCCTGTCCTGAAAGCCACAGACAATATGGTCCCAAATGACTG
ACTGCACCTTCTGTGCCTCAGCCGTTYTTGACATCAAGAATCTTCTGTTCCACA
TCCACACAGCCAATACAATTAGTCAAACCACTGTTATTAACAGATGTAGCAACA
TGAGAAACGCTTATGTTACAGGTTACATGAGAGCAATCATGTAAGTCTATATGA
CTTCAGAAATGTTAAAATAGACTAACCTCTAACAACAAATTAAAAGTGATTGTT
TCAAGGTGATGCAATTATTGATGACCTATTTTATTTTTCTATAATGATCATATA
TTACCTTTGTAATAAAACATTATAACCAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAA

VII-48 (Secuencia 15)

CTTAAGTATGCCCTGACAGGAGATGAAGTAAAGAAGATTTGCATGCAGCGGTTC
ATTAAAATCGATGGCAAGGTCCGAACTGATATAACCTACCCTGCTGGATTCATG
GATGTCATCAGCATTGACAAGACGGGAGAGAATTTCCGTCTGATCTATGACACC
AAGGGTCGCTTTGCTGTACATCGTATTACACCTGAGGAGGCCAAGTACAAGTTG
TGCAAAGTGAGAAAGATCTTTGTGGGCACAAAAGGAATCCCTCATCTGGTGACT
CATGATGCCCGCACCATCCGCTACCCCGATCCCCTCATCAAGGTGAATGATACC
ATTCAGATTGATTTAGAGACTGGCAAGATTACTGATTTCATCAAGTTCGACACT
GGTAACCTGTGTATGGTGACTGGAGGTGCTAACCTAGGAAGAATTGGTGTGATC
ACCAACAGAGAGAGGCACCCTGGATCTTTGACGTGGTTCACGTGAAAGATGCC
AATGGCAACAGCTTTGCCACTCGACTTTCCAACATTTTTGTTATTGGCAAGGGC
AACAAACCATGGATTTCTCTTCCCCGAGGAAAGGGTATCCGCCTCACCATTGCT
GAAGAGAGAGACAAAAGACTGGCGGCCAAACAGAGCAGTGGGTGAAATGGGTCC
CTGGGTGACATGTCAGATCTTTGTACGTAATTAAAAATATTGTGGCAGGATTAA
TAGCC

VII-76 (Secuencia 16)

AGACACACGAGCATATTTCACCTCCGCTACCATAATCATCGCTATCCCCACCGG
CGTCAAAGTATTTAGCTGACTCGCCACACTCCACGGAAGCAATATGAAATGATC
TGCTGCAGTGCTCTGAGCCCTAGGATTCATCTTTCTTTTCACCGTAGGTGGCCT
GACTGGCATTGTATTAGCAAACTCATCACTAGACATCGTACTACACGACACGTA
CTACGTTGTAGCCCACTTCCACTATGTCCTATCAATAGGAGCTGTATTTGCCAT
CATAGGAGGCTTCATTCACTGATTTCCCCTATTCTCAGGCTACACCCTAGACCA
AACCTACGCCAAAATCCATTTCACTATCATATTCATCGGCGTAAATCTAACTTT
CTTCCCACAACACTTTCTCGGCCTATCCGGAATGCCCCGACGTTACTCGGACTA
CCCCGATGCATACACCACATGAAACATCCTATCATCTGTAGGCTCATTCATTTC
TCTAACAGCAGTAATATTAATAATTTTCATGATTTGAGAAGCCTTCGCTTCGAA
GCGAAAAGTCCTAATAGTAGAAGAACCCTCCATAAACCTGGAGTGACTATATGG
ATGCCCCCCACCCTACCACACATTCGAAGAACCCGTATACAT

IX-24 (Secuencia 17)

AGAGTGCAAGACGATGACTTGCAAAATGTCGCAGCTGGAACGCAACATAGAGAC
CATCATCAACACCTTCCACCAATACTCTGTGAAGCTGGGGCACCCAGACACCCT
GAACCAGGGGGAATTCAAAGAGCTGGTGCGAAAAGATCTGCAAAATTTTCTCAA
GAAGGAGAATAAGAATGAAAAGGTCATAGAACACATCATGGAGGACCTGGACAC
AAATGCAGACAAGCAGCTGAGCTTCGAGGAGTTCATCATGCTGATGGCGAGGCT
AACCTGGGCCTCCCACGAGAAGATGCACGAGGGTGACGAGGGCCCTGGCCACCA

CCATAAGCCAGGCCTCGGGGAGGGCACCCCCTAAGACCACAGTGGCCAAGATCA
CAGTGGCCACGGCCACGGCCACAGTCATGGTGGCCACGGCCACAGCCACTAATC
AGGAGGCCAGGCCACCCTGCCTCTACCCAACCAGGGCCCCGGGGCCTGTTATGT
CAAACTGTCTTGGCTGTGGGGCTAGGGGCTGGGGCCAAATAAAGTCTCTTCCTC
CAAAAAAAA

IX-39 (Secuencia 18)

CTTGGCTCCTGTGGAGGCCTGCTGGGAACGGGACTTCTAAAAGGAACTATGTCT
GGAAGGCTGTGGTCCAAGGCCATTTTTGCTGGCTATAAGCGGGGTCTCCGGAAC
CAAAGGGAGCACACAGCTCTTCTTAAAATTGAAGGTGTTTACGCCCGAGATGAA
ACAGAATTCTATTTGGGCAAGAGATGCGCTTATGTATATAAAGCAAAGAACAAC
ACAGTCACTCCTGGCGGCAAACCAAACAAAACCAGAGTCATCTGGGGAAAAGTA
ACTCGGGCCCATGGAAACAGTGGCATGGTTCGTGCCAAATTCCGAAGCAATCTT
CCTGCTAAGGCCATTGGACACAGAATCCGAGTGATGCTGTACCCCTCAAGGATT
TAAACTAACGAAAAATCAATAAATAAATGTGGATTTGTGCTCTTGTA

IX-46 (Secuencia 19)

ACGCGAGATGGCAGTGCAAATATCCAAGAAGAGGAAGTTTGTCGCTGATGGCAT
CTTCAAAGCTGAACTGAATGAGTTTCTTACTCGGGAGCTGGCTGAAGATGGCTA
CTCTGGAGTTGAGGTGCGAGTTACACCAACCAGGACAGAAATCATTATCTTAGC
CACCAGAACACAGAATGTTCTTGGTGAGAAGGGCCGGCGGATTCGGGAACTGAC
TGCTGTAGTTCAGAAGAGGTTTGGCTTTCCAGAGGGCAGTGTAGAGCTTATGC
TGAAAAGGTGGCCACTAGAGGTCTGTGTGCCATTGCCCAGGCAGAGTCTCTGCG
TTACAAACTCCTAGGAGGGCTTGCTGTGCGGAGGGCCTGCTATGGTGTGCTGCG
GTTCATCATGGAGAGTGGGGCCAAAGGCTGCGAGGTTGTGGTGTCTGGGAAACT
CCGAGGACAGAGGGCTAAATCCATGAAGTTTGTGGATGGCCTGATGATCCACAG
CGGAGACCCTGTTAACTACTACGTTGACACTGCTGTGCGCCACGTGTTGCTCAG
ACAGGGTGTGCTGGGCATCAAGGTGAAGATCATGCTGCCCTGGGACCCAACTGG
TAAGATTGGCCCTAAGAAGCCCCTGCCTGACCACGTGAGCATTGTGGAACCCAA
AGATGAGATACTGCCCACCACCCCCATCTCAGAACAGAAGGGTGGGAAGCCAGA
GCCGCCTGCCATGCCCCAGCCAGTCCCCACAGCATAACAGGGTCTCCTTGGCAG
CTGTATTCTGGAGTCTGGATGTTGCTCTCTAAAGACCTTTAATAAAATTTTGT

IX-50 (Secuencia 20)

GTCCATCCTGCAGGCCACAAGCTCTGGATGAGGAACTTGAGGCAAGTCACCAGC
CCCTGATCATTTCGCCTAAAAGAGCAAGGACTAGAGTTCCTGACCTCCAGGCCA
GTCCCTGATCCCTGACCTAATGTTATCGCGGAATGATGATATATGTATCTACGG

GGGCCTGGGGCTGGGCGGGCTCCTGCTTCTGGCAGTGGTCCTTCTGTCCGCCTG
CCTGTGTTGGCTGCATCGAAGAGTAAAGAGGCTGGAGAGGAGCTGGGCCCAGGG
CTCCTCAGAGCAGGAACTCCACTATGCATCTCTGCAGAGGCTGCCAGTGCCCAG
CAGTGAGGGACCTGACCTCAGGGGCAGAGACAAGAGAGGCACCAAGGAGGATCC
AAGAGCTGACTATGCCTGCATTGCTGAGAACAAACCCACCTGAGCACCCCAGAC
ACCTTCCTCAACCCAGGCGGGTGGACAGGGTCCCCCTGTGGTCCAGCCAGTAAA
AACCATGGTCCCCCCACTTCTGTGTCTCAGTCCTCTCAGTCCATCTCGAGCCTC
CGTTCAAAATGATCATCATCAAAACTTATGTGGCTTTTTGACCTTTGAATAGGG
AATTTTTTAAATTTTTTAAAAATTAAAATAAAAAAAACACATGGCTCACCCTTC
CACCCAAAAAAAAAA

X-77 (Secuencia 21)

CCTCCCGGGCTCTTAAGCCCCTCTCTTTCTCTAACAGAAAAAGCGGATGGTGGT
TCCTGCTGCCCTCAAGGTCGTGCGTCTGAAGCCTACAAGAAAGTTTGCCTATCT
GGGGCGCCTGGCTCACGAGGTTGGCTGGAAGTACCAGGCAGTGACAGCCACCCT
GGAGGAGAAGAGGAAAGAGAAAGCCAAGATCCACTACCGGAAGAAGAAACAGCT
CATGAGGCTACGGAAACAGGCCGAGAAGAACGTGGAGAAGAAAATTGACAAATA
CACAGAGGTCCTCAAGACCCACGGACTCCTGGTCTGAGCCCAATAAAGACTGTT
AATTCCTCATGCGTTGCCTGCCCTTCCTCCATTGTTGCCCTGGAATGTACGGGA
CCCAGGGGCAGCAGCAGTCCAGGTGCCACAGGCAGCCCTGGGACATAGGAAGCT
GGGAGCAAGGAAAGGGTCTTAGTCACTGCCTCCCGAAGTTGCTTGAAAGCACTC
GGAGAATTGTGCAGGTGTCATTTATCTATGACCAATAGGAAGAGCAACCAGTTA
CTATGAGTGAAAGGGAGCCAGAAGACTGATTGGAGGGCCCTATCTTGTGAGTGG
GGCATCTGTTGGACTTTCCACCTGGTCATATACTCTGCAGCTGTTAGAATGTGC
AAGCACTTGGGGACAGCATGAGCTTGCTGTTGTACACAGGGTATT

XI-13 (Secuencia 22)

CTGCCAACATGGTGTTCAGGCGCTTCGTGGAGGTTGGCCGGGTGGCCTATGTCT
CCTTTGGACCTCATGCCGGAAAATTGGTCGCGATTGTAGATGTTATTGATCAGA
ACAGGGCTTTGGTCGATGGACCTTGCACTCAAGTGAGGAGACAGGCCATGCCTT
TCAAGTGCATGCAGCTCACTGATTTCATCCTCAAGTTTCCGCACAGTGCCCACC
AGAAGTATGTCCGACAAGCCTGGCAGAAGGCAGACATCAATACAAAATGGGCAG
CCACACGATGGGCCAAGAAGATTGAAGCCAGAGAAAGGAAAGCCAAGATGACAG
ATTTTGATCGTTTTAAAGTTATGAAGGCAAAGAAAATGAGGAACAGAATAATCA
AGAATGAAGTTAAGAAGCTTCAAAAGGCAGCTCTCCTGAAAGCTTCTCCCAAAA
AAGCACCTGGTACTAAGGGTACTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
CTGCTGCTGCTGCTGCTAAAGTTCCAGCAAAAAAGATCACCGCCGCGAGTAAAA

AGGCTCCAGCCCAGAAGGTTCCTGCCCAGAAAGCCACAGGCCAGAAAGCAGCGC
CTGCTCCAAAAGCTCAGAAGGGTCAAAAAGCTCCAGCCCAGAAAGCACCTGCTC
CAAAGGCATCTGGCAAGAAAGCATAAGTGGCAATCATAAAAAGTAATAAAGGTT
CTTTTTGACCTGTTAAAAAA

XI-49 (Secuencia 23)

GATCAACCTGGAGCTCTACGCCTCCTACGTTTACCTGTCCATGTCTTACTACTT
TGACCGCGATGATGTGGCTTTGAAGAACTTTGCCAAATACTTTCTTCACCAATC
TCATGAGGAGAGGGAACATGCTGAGAAACTGATGAAGCTGCAGAACCAACGAGG
TGGCCGAATCTTCCTTCAGGATATCAAGAAACCAGACTGTGATGACTGGGAGAG
CGGGCTGAATGCAATGGAGTGTGCATTACATTTGGAAAAAAATGTGAATCAGTC
ACTACTGGAACTGCACAAACTGGCCACTGACAAAAATGACCCCCATTTGTGTGA
CTTCATTGAGACACATTACCTGAATGAGCAGGTGAAAGCCATCAAAGAATTGGG
TGACCACGTGACCAACTTGCGCAAGATGGGAGCGCCCGAATCTGGCTTGGCGGA
ATATCTCTTTGACAAGCACACCCTGGGAGACAGTGATAATGAAAGCTAAGCCTC
GGGCTAATTTCCCCATAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCA
TGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCCACTT
AAGTTCTTTGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACCC

XI-81 (Secuencia 24)

AGAGCAGCAGCCATGGCCCTACGCTACCCTATGGCCGTGGGCCTCAACAAGGGC
CACAAAGTGACCAAGAACGTGAGCAAGCCCAGGCACAGCCGACGCCGCGGGCGT
CTGACCAAACACACCAAGTTCGTGCGGGACATGATTCGGGAGGTGTGTGGCTTT
GCCCCGTACGAGCGGCGCGCCATGGAGTTACTGAAGGTCTCCAAGGACAAACGG
GCCCTCAAATTTATCAAGAAAAGGGTGGGGACGCACATCCGCGCCAAGAGGAAG
CGGGAGGAGCTGAGCAACGTACTGGCCGCCATGAGGAAAGCTGCTGCCAAGAAA
GACTGAGCCCCTCCCCTGCCCTCTCCCTGAAATAAA

XII-35 (Secuencia 25)

CTCTCCTGTCAACAGCGGCCAGCCTCCCAACTACGAGATGCTCAAGGAGGAGCA
GGAAGTGGCTATGCTGGGGGCGCCCCACAACCCTGCTCCCCCGACGTCCACCGT
GATCCACATCCGCAGCGAGACCTCCGTGCCCGACCATGTCGTCTGGTCCCTGTT
CAACACCCTCTTCATGAACACCTGCTGCCTGGGCTTCATAGCATTCGCCTACTC
CGTGAAGTCTAGGGACAGGAAGATGGTTGGCGACGTGACCGGGGCCCAGGCCTA
TGCCTCCACCGCCAAGTGCCTGAACATCTGGGCCCTGATTTTGGGCATCTTCAT
GACCATTCTGCTCGTCATCATCCCAGTGTTGGTCGTCCAGGCCCAGCGATAGAT
CAGGAGGCATCATTGAGGCCAGGAGCTCTGCCCGTGACCTGTATCCCACGTACT

CTATCTTCCATTCCTCGCCCTGCCCCCAGAGGCCAGGAGCTCTGCCCTTGACCT
GTATTCCACTTACTCCACCTTCCATTCCTCGCCCTGTCCCCACAGCCGAGTCCT
GCATCAGCCCTTTATCCTCACACGCTTTTCTACAATGGCATTCAATAAAGTGTA
TATGTTTCTGGTGCTGCTGTGACTTCAA

XII-77 (Secuencia 26)

GTAAGAAAGCCCTTAAATAAAGAAGGTAAGAAACCTAGGACCAAAGCACCCAAG
ATTCAGCGTCTTGTTACTCCACGTGTCCTGCAGCACAAACGGCGGCGTATTGCT
CTGAAGAAGCAGCGTACCAAGAAAAATAAAGAAGAGGCTGCAGAATATGCTAAA
CTTTTGGCCAAGAGAATGAAGGAGGCTAAGGAGAAGCGCCAGGAACAAATTGCG
AAGAGACGCAGACTTTCCTCTCTGCGAGCTTCTACTTCTAAGTCTGAATCCAGT
CAGAAATAAGATTTTTTGAGTAACAAATAAATAAGATCAGACTCTGAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

XIII-29 (Secuencia 27)

CTGGCTCACGCAGCACTCGTGGCAGTCCCTGAAGGACCGCTACCTCAAGCACCT
GCGGGGCCAGGAGCATAAGTACCTGCTGGGGGACGCGCCGGTGAGCCCCTCCTC
CCAGAAGCTCAAGCGGAAGGCGGAGGAGGACCCGGAGGCCGCGGATAGCGGGGA
ACCACAGAATAAGAGAACTCCAGATTTGCCTGAAGAAGAGTATGTGAAGGAAGA
AATCCAGGAGAATGAAGAAGCAGTCAAAAAGATGCTTGTGGAAGCCACCCGGGA
GTTTGAGGAGGTTGTGGTGGATGAGAGCCCTCCTGATTTTGAAATACATATAAC
TATGTGTGATGATGATCCACCCACACCTGAGGAAGACTCAGAAACACAGCCTGA
TGAGGAGGAAGAAGAAGAAGAAGAAAAAGTTTCTCAACCAGAGGTGGGAGCTGC
CATTAAGATCATTCGGCAGTTAATGGAGAAGTTTAACTTGGATCTATCAACAGT
TACACAGGCCTTCCTAAAAAATAGTGGTGAGCTGGAGGCTACTTCCGCCTTCTT
AGCGTCTGGTCAGAGAGCTGATGGATATCCCATTTGGTCCCGACAAGATGACAT
AGATTTGCAAAAAGATGATGAGGATACCAGAGAGGCATTGGTCAAAAAATTTGG
TGCTCAGAATGTAGCTCGGAGGATTGAATTTCGAAAGAAATAATTGGCAAGATA
ATGAGAAAAGAAAAAAGTCATGGTAGGTGAGGTGGTTAAAAAAAATTGTGACCA
ATGAACTTTAGAGAGTTCTTGCATTGGAACTGGCACTTATTTTCTGACCATCGC
TGCTGTTGCTCTGTGAGTCCTAGATT

XIII-84 (Secuencia 28)

ATTATCCTCAGTTCCCAAGAGCAATCATACTTTTCCACACATACCGTGTGTCTC
ATGTTAGGTAAATGTATTTTTACAATGAGCACCACTTCTGTGGAAAAAGTTCCC
TGCACGGGGAGGTCCAGCTTCCAGACTGCTCCATCGCATAAGGACTTCCCCATT
CCCCTAAATGCTGCTCTGTCAGAACCTGCCCAGGTAATGGTAATGACCCTAGAG

AGATGATTTCTGAACCGCAATTTTGAGCCCATTAGAAGGTGTGTGGTGGGCATT
TATTTCATCCTGATGCTCTGGTGAGAATCTTTGCAGACGCACTAGATCCAGAAG
CTGTTAATCTTGGTGCATTTATTTTCCTACCTAAAAGAACCAAGCAGCTCAGAG
GCAGTGACTGTACAGGATGCAGTGTTTATAATAATGCTGAGCTTGCTGGTCTGG
AACCCCACACTTCAGCAATCCCAGCATTGTTCCTGTTTATGAAGTTGACAAAGT
GACCAGGGCAAGGGGGTATTATCATTAAATACACTCTAGGAGAGGCAGAACACA
TGAGGGCAATGTTTTTCAGAGGTCTTTAGGCCACCGCATCAGATTCTCCTGGAG
CATAAAGCAAATGCTTTATGAGTCCAGGGCCCCTGCAGACCTACTGTATACTAG
TATACAGCTCCCTCTTAGTGGATCTCAAGCTTGTTTCCAAAAAGTCATTACACT
CCTTACCAAAGCCCATGACACATTCATACAGATTCATCCAGACATAACCCACTG
CATGGTCCAGTGCATGCTTGTGTGCTTAACTTATTATAGATCAAGTGTTATTTA
AGTCCAACATATTAAACGTGACTGAATATT

XV-49 (Secuencia 29)

AAGTCTGCCCAGAAAGCTCAGAAGGCTAAATGAATATTATCCCTAATACCTGCC
ACCCCACTCTTAATCAGTGGTGGAAGAACGGTCTCAGAACTGTTTGTTTCAATT
GGCCATTTAAGTTTAGTAGTAAAAGACTGGTTAATGATAACAATGCATCGTAAA
ACCTTCAGAAGGAAAGGAGAATGTTTTGTGGACCACTTTGGTTTTCTTTTTTGC
GTGTGGCAGTTTTAAGTTATTAGTTTTTAAAATCAGTACTTTTTAATGGAAACA
ACTTGACCAAAAATTTGTCACAGAATTTTGAGACCCATTAAAAAAGTTAAATGAG

XV-54 (Secuencia 30)

AAGAGCAGGTCTCTGGAGGCTGAGTTGCATGGGGCCTAGTAACACCAAGCCAGT
GAGCCTCTAATGCTACTGCGCCCTGGGGGCTCCCAGGGCCTGGGCAACTTAGCT
GCAACTGGCAAAGGAGAAGGGTAGTTTGAGGTGTGACACCAGTTTGCTCCAGAA
AGTTTAAGGGGTCTGTTTCTCATCTCCATGGACATCTTCAACAGCTTCACCTGA
CAACGACTGTTCCTATGAAGAAGCCACTTGTGTTTTAAGCAGAGGCAACCTCTC
TCTTCTCCTCTGTTTCGTGAAGGCAGGGGACACAGATGGGAGAGATTGAGCCAA
GTCAGCCTTCTGTTGGTTAATATGGTATAATGCATGGCTTGTGCACAGCCCAG
TGTGGGATTACAGCTTTGGGATGACCGCTTACAAAGTTCTGTTTGGTTAGTATT
GGCATAGTTTTTCTATATAGCCATAAATGCGTATATATACCCATAGGGCTAGAT
CTGTATCTTAGTGTAGCGATGTATACATATACACATCCACCTACATGTTGAAGG
GCCTAACCAGCCTTGGGAGTATTGACTGGTCCCTTACCTCTTATGGCTAAGTCT
TTGACTGTGTTCATTTACCAAGTTGACCCAGTTTGTCTTTTAGGTTAAGTAAGA
CTCGAGAGTAAAGGCAAGGAGGGGGGCCAGCCTCTGAATGCGGCCACGGATGCC
TTGCTGCTGCAACCCTTTCCCCAGCTGTCCACTGAAACGTGAAGTCCTGTTTTG
AATGCCAAACCCACCATTCACTGGTGCTGACTACATAGAATGGGGTTGAGAGAA
GATCAGTTTGGGCTTCACAGTGTCATTTGAAAACGTTTTTTGTTTTGTTTTGTA

ATTATTGTGGAAAACTTTCAAGTGAACAGAAGGATGGTGTCCTACTGTGGATGA
GGGATGAACAAGGGGATGGCTTTGATCCAATGGAGCCTGGGAGGTGTGCCCAGA
AAGCTTGTCTGTAGCGGGTTTTGTGAGAGTGAACACTTTCCACTTTTTGACACC
TTATCCTGATGTATGGTTCCAGGATTTGGATTTTGATTTTCCAAATGTAGCTTG
AAATTTCAATAAACTTTGCTCTGTTTTTCTAAAAATAAAAAAAAAAAAAAAAAA
AAAAAAAA

XV-75 (Secuencia 31)

AGCAGATGACCCTTCGTGGCACCCTCAAGGGCCACAACGGCTGGGTAACCCAGA
TCGCTACTACCCCGCAGTTCCCGGACATGATCCTCTCCGCCTCTCGAGATAAGA
CCATCATCATGTGGAAACTGACCAGGGATGAGACCAACTATGGAATTCCACAGC
GTGCTCTGCGGGGTCACTCCCACTTGTTAGTGATGTGGTTATCTCCTCAGATG
GCCAGTTTGCCCTCTCAGGCTCCTGGGATGGAACCCTGCGCCTCTGGGATCTCA
CAACGGGCACCACCACGAGGCGATTTGTGGGCCATACCAAGGATGTGCTGAGTG
TGGCCTTCTCCTCTGACAACCGGCAGATTGTCTCTGGATCTCGAGATAAAACCA
TCAAGCTATGGAATACCCTGGGTGTGTGCAAATACACTGTCCAGGATGAGAGCC
ACTCAGAGTGGGTGTCTTGTGTCCGCTTCTCGCCCAACAGCAGCAACCCTATCA
TCGTCTCCTGTGGCTGGGACAAGCTGGTCAAGGTATGGAACCTGGCTAACTGCA
AGCTGAAGACCAACCACATTGGCCACACAGGCTATCTGAACACGGTGACTGTCT
CTCCAGATGGATCCCTCTGTGCTTCTGGAGGCAAGGATGGCCAGGCCATGTTAT
GGGATCTCAACGAAGGCAAACACCTTTACACGCTAGATGGTGGGGACATCATCA
ACGCCCTGTGCTTCAGCCCTAACCGCTACTGGCTGTGTGCTGCCACAGGCCCCA
GCATCAAGATCTGGGATTTAGAGGGAAAGATCATTGTAGATGAACTGAAGCAAG
AAGTTATCAGTACCAGCAGCAAGGCAGAACCACCCCAGTGCACCTCCCTGGCCT
GGTCTGCTGATGGCCAGACTCTGTTTGCTGGCTACACGGACAACCTGGTGCGAG
TGTGGCAGGTGACCATTGGCACACGCTAGAAGTTTATGGCAGAGCTTTACAAAT
AAAAAAAAAACTGGCTTTTCTGACAAAAAAAAAA

XV-86 (Secuencia 32)

GCAAAATGTCGCAGCTGGAACGCAACATAGAGACCATCATCAACACCTTCCACC
AATACTCTGTGAAGCTGGGGCACCCAGACACCCTGAACCAGGGGGAATTCAAAG
AGCTGGTGCGAAAAGATCTGCAAAATTTTCTCAAGAAGGAGAATAAGAATGAAA
AGGTCATAGAACACATCATGGAGGACCTGGACACAAATGCAGACAAGCAGCTGA
GCTTCGAGGAGTTCATCATGCTGATGGCGAGGCTAACCTGGGCCTCCCACGAGA
AGATGCACGAGGGTGACGAGGGCCCTGGCCACCACCATAAGCCAGGCCTCGGGG
AGGGCACCCCCTAAGACCACAGTGGCCAAGATCACAGTGGCCACGGCCACGGCC

ACAGTCATGGTGGCCACGGCCACAGCCACTAATCAGGAGGCCAGGCCACCCTGCCT
CTACCCAACCAGGGCCCCGGGGCCTGTTATGTCAAACTGTCTTGGCTGTGGG
GCTAGGGGCTGGGGCCAAATAAAGTCTCTTCCTCCAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAA

XVI-74 (Secuencia 33)

CGCCGCCGCGCCGCCGTCGCTCTCCAACGCCAGCGCCGCCTCTCGCTCGCCGAG
CTCCAGCCGAAGGAGAAGGGGGGTAAGTAAGGAGGTCTCTGTACCATGGCTCGT
ACAAAGCAGACTGCCCGCAAATCGACCGGTGGTAAAGCACCCAGGAAGCAACTG
GCTACAAAAGCCGCTCGCAAGAGTGCGCCCTCTACTGGAGGGGTGAAGAAACCT
CATCGTTACAGGCCTGGTACTGTGGCGCTCCGTGAAATTAGACGTTATCAGAAG
TCCACTGAACTTCTGATTCGCAAACTTCCCTTCCAGCGTCTGGTGCGAGAAATT
GCTCAGGACTTTAAAACAGATCTGCGCTTCCAGAGCGCAGCTATCGGTGCTTTG
CAGGAGGCAAGTGAGGCCTATCTGGTTGGCCTTTTTGAAGACACCAACCTGTGT
GCTATCCATGCCAAACGTGTAACAATTATGCCAAAAGACATCCAGCTAGCACGC
CGCATACGTGGAGAACGTGCTTAAGAATCCACTATGATGGGAAACATTTCATTC
TCAAAAAAAAAAAAAAAAAATTTCTCTTCTTCCTGTTATTGGTAGTTCTGAACG
TTAGATATTTTTTTTCCATGGGGTCAAAAGGTACCTAAGTATATGATTGCGAGT
GGAAAAATAGGGGACAGAAATCAGGTATTGGCAGTTTTTCCATTTTCATTTGTG
TGTGAATTTTTAATATAAATGCGGAGACGTAAAGCATTAATGCAAGTTAAAATG
TTTCAGTGAACAAGTTTCAGCGGTTCAACTTTATAATAATTATAAATAAACCTG
TTAAATTTTTCTGGACAATGCCAGCATTTGGATTTTTTTAAAACAAGTAAATTT
CTTATTGATGGCAACTAAATGGTGTTTGTAGCATTTTTATCATACAGTAGATTC
CATCCATTCACTATACTTTTCTAACTGAGTTGTCCTACATGCAAGTACATGTTT
TTAATGTTGTCTGTCTTCTGTGCTGTTCCTGTAAGTTTGCTATTAAAATACATT
AAACTATAAAAAAAAAAAAAAAAAA

XVII-77 (Secuencia 34)

CAGACACCCTGAACCAGGGGGAATTCAAAGAGCTGGTGCGAAAAGATCTGCAAA
ATTTTCTCAAGAAGGAGAATAAGAATGAAAAGGTCATAGAACACATCATGGAGG
ACCTGGACACAAATGCAGACAAGCAGCTGAGCTTCGAGGAGTTCATCATGCTGA
TGGCGAGGCTAACCTGGGCCTCCCACGAGAAGATGCACGAGGGTGACGAGGGCC
CTGGCCACCACCATAAGCCAGGCCTCGGGGAGGGCACCCCCTAAGACCACAGTG
GCCAAGATCACAGTGGCCACGGCCACGGCCACAGTCATGGTGGCCACGGCCACA
GCCACTAATCAGGAGGCCAGGCCACCCTGCCTCTACCCAACCAGGGCCCCGGGG
CCTGTTATGTCAAACTGTCTTGGCTGTGGGGCTAGGGGCTGGGGCCAAATAAAG
TCTCTTCCTCCAAAAAAA

XII-78 sin secuencia disponible

Tabla 1. Detalles de las muestras. Fase 0, carcinoma *in situ*; Fase I, carcinoma invasivo con tamaño de tumor < 20 mm; Fase II, carcinoma invasivo con tamaño de tumor > 20-50 mm; Fase III, carcinoma invasivo con tamaño de tumor > 50 mm; Fase IV, cáncer diseminado a partes alejadas. IDC, carcinoma ductal invasivo (del inglés, invasive ductal carcinoma); DCIS, carcinoma ductal *in situ*; ILC, carcinoma lobular invasivo (del inglés, invasive lobular carcinoma). nd, no disponible. SD, sin decisión. *Muestras sanguíneas tomadas de la misma hembra en cinco semanas consecutivas

Id. de la hembra	*Edad*	*Fase*	*Histología*	*Grado*	*Tamaño (mm)*	*Nódulos*	*Otra enfermedad, si la hubiera/comentarios*	*Nº de análisis*	*Predicción final*
1	51	II	IDC	3	20	1/7	–	2	+
2	84	II	IDC	1	22	2/2	–	2	+
3	50	I	IDC (multifocal)	1	5 x 14	0	–	1	+
4	66	I	IDC	2	15	0	Enfermedad reumática	3	+
5	66	II	IDC	1	26	0	Epilepsia	1	+
6	47	I	IDC	2	15	0	–	2	SD
7	69	III	ILC + adenocarcinoma tubular	2+1	50 + 3	2/19	–	2	SD
8	50	II	IDC	2	24	0	–	2	+
9	65	I	IDC	1	15	0	–	1	–
10	63	II	IDC	3	23	0	–	1	+
11	65	IV				Metástasis en nódulos supra- e infraclaviculares	–	1	–
12	52	I	IDC	1	3	0	–	2	+
13	60	II	IDC	2	23	0	–	2	+
14	54	I	IDC	1	11	0	–	2	+
15	67	0	DCIS	2	20	0	–	3	+
16	nd	0	DCIS	2	9	0	–	1	–
17	48	I	IDC	2	4	0	–	2	+
18	nd	I	IDC	2	14	0	Psoriasis	1	+
19	68	I	IDC	1	7	0	–	1	+
20	63	I	IDC	1	10	0	–	2	+
21	65	I	IDC	1	11	0	Diabetes tipo II	3	+
22	44	II	IDC	2	25	0	–	1	+
23	55	III	IDC	1	35	0	–	1	+
24	71	I	IDC	1	8	0	–	1	+

Subgrupo A2: Mujeres con primera mamografía anormal

Id. de la hembra	*Edad*	*Anormalidad de mama*	*Otra enfermedad, si la hubiera/comentarios*	*Nº de análisis*	*Predicción final*
25	44	Densidad benigna	–	2	+
26	46	Densidad benigna	–	2	+
27	53	Microcalcificaciones benignas	Quiste encapsulado en rodilla izquierda	2	+
28	52	Microcalcificaciones benignas	Cáncer, intestino grueso, 1992	1	SD
29	45	Densidad benigna	–	2	+
30	59	Tumor benigno, fibroadenoma	–	2	+
31	46	Densidad benigna	–	2	+
32	46	Densidad benigna	Colitis ulcerosa desde 1983	2	SD
33	50	Densidad benigna	Diabetes de tipo I	2	+
34	47	Microcalcificaciones benignas	–	2	+
35	46	Densidad benigna, quiste	Enfermedad de Crohn	2	+
36	nd	Densidad benigna	Enfermedad reumática	1	+
37	44	Microcalcificaciones benignas	–	2	+
38	47	Densidad benigna	–	2	+
39	50	Fibrosis, benigna	Histología, 60 mm de tamaño	1	+
40	45	Densidad benigna	Diabetes de tipo II	2	+
41	63	Densidad benigna, quiste	Fibromialgia	2	+
42	44	Densidad benigna	–	2	+
43	51	Cicatriz radial	Histología, 10 mm de tamaño	1	+

Subgrupo A3: Mujeres sin anormalidad en la mama

Id. de la hembra	*Edad*	*Comentarios*	*Nº de análisis*	*Predicción*
44	22	–	2	+
45	34	Encinta, 8 meses	3	+
46	27	Encinta, 6 meses	1	+
47*	18	Semana 1	2	+
		Semana 2	1	+
		Semana 3	1	+
		Semana 4	2	+
		Semana 1	1	+
48	29	Encinta, 9 meses	1	–
49	30	Amamantamiento	2	+
50	26	–	1	+
51	43	–	1	+
52	42	–	3	+
53	43	–	2	+
54	34	Amamantamiento	3	+
55	–	–	1	+
56	51	Infección bacteriana aguda además de infección crónica por EBV	1	+

Tabla 2. Detalles de los 35 genes significativos seleccionados mediante Jackknife. Se muestran su posición en la red, la identificación del clon así como el número de acceso de secuencias de bases de datos públicas que se corresponden con ellos, y su conocida o supuesta función celular

Genes suprarregulados

Id. del clon	*Id. de la posición*	*Nº de acceso*	*Similitud génica*	*Supuesta función biológica*	*Nº de secuencia*
III-2	6A	sin éxito	–	–	2
III-27	10M	AC096970	Clon RP11-321A23 del cromosoma 3; procedente de la secuencia nº 135183-135446	–	3
III-60	14AC	NM_001665	Familia génica de homólogos de Ras, miembro G (rho-G)	Transducción de señales ¿inhibidor de cinasas? (RhoH ha sido descrito como un inhibidor de cinasas)	4
IV-26	6N	BC016857	Ferritina, polipéptido pesado 1	Almacenamiento de hierro; defensa contra ROS	5
IV-51	10Z	BC042655	Factor 2 de transcripción cadena arriba, USF2	Regulador de la transcripción	7
VI-44	14X	AC087441	Cromosoma 11; procedente de la secuencia nº 116068-116692	–	10
VI-52	14AB	sin éxito	–	–	12
VII-15	27E	BC001431	Proteína A6 (calciclina) de S100, ligante de calcio	Defensa; inhibición de la caseína cinasa II	13
VII-32	31M	M28697	Receptor humano de baja afinidad para Fc de IgG (alfa-Fc-gamma-RII)	Respuesta inmune	14
IX-24	31J	BC047681	Proteína A9 (calgranulina B) de S100, ligante de calcio	Defensa; inhibición de la caseína cinasa II	17
IX-50	7Z	NM_007161	Transcrito 1 específico de leucocitos	Relacionado con la defensa	20
XI-49	3AB	BC016857	Ferritina, polipéptido pesado 1, mRNA	Almacenamiento de hierro; defensa contra ROS	23
XII-35	12Q	BC009696	Proteína transmembranal 2 inducida por interferón	Respuesta inmune	25
XII-78	24K	–	–	–	–
XIII-84	16AP	AL391903	De la secuencia nº 75875-76710	–	28
XV-54	24AA	BC018148	Péptido inductor del sueño delta, inmunorreactivo	¿Respuesta inmune?	30
XV-86	24AQ	BC047681	Proteína A9 (calgranulina B) de S100, ligante de calcio	Defensa; inhibición de la caseína cinasa II	32
XVI-74	5AK	BC066901	Histona H3, familia 3B (H3.3B)	Remodelación de la cromatina	33
XVII-77	20AN	BC047681	Proteína A9 (calgranulina B) de S100, ligante de calcio	Defensa; inhibición de la caseína cinasa II	34

Tabla 2 (continuación)

Genes infrarregulados

Id. del clon	*Id. de la posición*	*Nº de acceso*	*Similitud génica*	*Supuesta función biológica*	*Nº de secuencia*
I-30	21O	BC009689	Proteína ligante de tipo ciclina D	Transcripción mediada por E2F	1
IV-41	2V	BC010165	Proteína ribosómica S2	Producción de ribosomas	6
V-09	2G	BC053370	Proteína ribosómica SA	Producción de ribosomas	8
V-38	22S	NM_001016	Proteína ribosómica S12	Producción de ribosomas	9
VI-49	2AB	NM_001023	Proteína ribosómica S20 (RPS20)		11
VII-48	31U	M22146	Proteína ribosómica S4	Producción de ribosomas	15
VII-76	15AK	AY495316	Subunidad de la citocromo c oxidasa, COX 1	Cadena de transporte electrónico mitocondrial	16
IX-39	27R	BC001037	Proteína ribosómica L35a	Producción de ribosomas	18
IX-46	23V	BC034149	Proteína ribosómica S3	Producción de ribosomas	19
X-77	19AM	BC000514	Proteína ribosómica L13a	Producción de ribosomas	21
XI-13	19H	D87735	Proteína ribosómica L14	Producción de ribosomas	22
XI-81	3AR	AF077043	Proteína ribosómica L36 de 60S	Producción de ribosomas	24
XII-77	20AK	BC035447	Proteína ribosómica S6	Producción de ribosomas	26
XIII-29	20N	BC004465	Factor 2 ligante de repeticiones teloméricas, proteína interactiva	Regulación de la longitud de telómeros	27
XV-49	4AA	BC018641	Factor 1 alfa 1 eucariótico de elongación de la traducción (EEF1A)	Traducción de proteínas	29
XV-75	12AM	BC019093	Proteína ligante de nucleótidos de guanina, tipo polipéptido beta 2; RACKs (para receptores para cinasa C activada)	Traducción de proteínas	31

Tabla 3: Sondas informativas para cáncer de mama – genes de las familias (i) y (ii)

Nº de sonda	Id. Agilent	Secuencia oligonucleotídica
2	A_23_P164011	ACTCCAGACTGGGAAGACCTTTCCATTTTCAGGATCGACGCTTCACGTTGAGGGGAGGGC
3	A_23_P94111	TTACCAAACTCAAAGCTTATTTGAGTAGAATGGGCTCATGGGCAATGTGATGTTCCCTGT
5	A_23_P155009	TGTTGGTTGGAGGACAAGTGGGCACTGAGACCCTGGTGACCCATGGAAAGGGTGGGCCTG
6	A_23_P84323	TGGAGAAAGGACCCTGGACCTGTGGGTCCATCGTCCGTTCCAGGAGCAGGCAGGCTGGGG
8	A_23_P121716	TGGACATTCGAACAGAGTTCAAGAAGCATTATGGCTATTCCCTATATTCAGCAATTAAAT
10	A_23_P111037	ATCAGAAGTCCACTGAACTGCTTATTCGTAAACTACCTTTCCAGCGCCTGGTGCGCGAGA
13	A_23_P76830	TTTGTGGAAACTGTGTGTTATACTTTGTGGTATAGACTGCCTGTTTAGTATGAAGGGGCG
16	A_23_P149936	CCTCCCAGCAGTTAAGTAACTTGTGTGAAGATGGGACCCTTGTTCCTAATGGTTCTAGAA
17	A_23_P134805	CTGAATCTGTTTTGTCTTCCTAATCTATCACAATTGCCACCCATCGGGTTTTGGGTGTGT
18	A_23_P154235	CCATGTTTCTGAATCTTCTTTGTTTCAAATGGTGCTGCATGTTTTCAACTACAATAAGTG
19	A_23_P2616	ATCATTCAGAATCTGAAAAGAAATTCTTCTTATTTTCTGGGGCTGTCAGATCCAGGGGGT
20	A_23_P333484	CCACCGAGCTGCTGATCAGAAAGCTGCCTTTTCAGCGTCTGGTGCGTGAGATCGCGCAGG
24	A_23_P259874	TCTCAGAAGAATGTTGGCCATGAGACTATCATTCAGAGGAGGAGGGGATTTCTCTCTTCA
26	A_23_P206568	AATCCTGTGATTCTGTGTGTGCCTGTGTGTGTATGCTGTTAATAAGATAAGGCTGCCCAT
27	A_23_P115091	GATGGCTGAAGGAGCTCTATGACCATGCTGAAGCCACGATCGTCGTCATGCTCGTGGGTA
28	A_23_P46718	TGCATGGGGAGTACATTCATCTGGAGGCTGCGTCCTGATGAATGTCCTGTCTGCTGGGGT
29	A_23_P218456	GTTTTTGAGTTTTTGCAGTTCAGTATCCCTCTGTCTATTCACACTTCGTGTTAGTGGTAA
31	A_23_P76610	CAGTTTATGGATGTCTGGGCAATCATAGCACTTGCCATTTAAAAACATGCTACAGGGGCA
32	A_23_P206396	ATTATCAACTCACTGGTAACAACAGTATTCATGCTCATCGTATCTGTGTTGGCACTGATA
34	A_23_P56091	GAAACCGGATCGCAAGCTTCCCAGGATTCCTCTTCGTGCTGCTGGGGGTGGGAAGCATGG
35	A_23_P55184	TCAATTTCAAGGCCTCCCTGCCTCTACTAGGCGCCTTAGCTCACTATGGGGAACCACTTG
37	A_23_P150974	CAAAATAGCTACATCCCTGAACACAGTCCGGAATATTACGGCCGGACCAGGGAATCGGGA
40	A_23_P111689	TTAATTCTATTGGCTCTTAGTCACTTGGAACTGATTAATTCTGACTTTCTGTCACTAAGC
41	A_23_P58937	GTCTCAAACAGCCGAAACCTGTCTTGCAATGGGGGGAGGGGGCGTTTCGCTTTCCTTCTT
42	A_23_P74828	TTGGCTTTTAGACATTATATATATTATCAGAGAAGTAGCCTAGTGGTCGTGGGGCACAGA
45	A_23_P42168	GGAACACTGTGAAAGTTACTTGGGGAGGGTGGGCCGGTGGGGCCGTAGCTCTCTACCTCT
47	A_23_P81278	TCAGACAGAGCTTGGTAAGTGACCCCTCTTAGAACTATTTCTCCTCAGGGCCGGGTCCAG

ES 2 358 910 T3

49	A_23_P251695	AGGTTGAACTCTTTTTTGTTGCTCAAGTTCTAGGAGTCCCTTTCCTGAATATATACTTGT
52	A_23_P393645	CTACTTTAGAGTCTTCTCCAATGTCCAAAAGGCTAGGGGGTTGGAGGTGGGGACTCTGGA
53	A_23_P206683	ATAGTCATGGGTGTCATGAAAAAATACCAAATGTAAGAGAACCTCCAAGTCAGGGCGCAG
55	A_23_P72016	GACATTGAGAAGGAAAACCGGGAGGTGGGAGACTGGCGCAAGAATATCGATGCACTAAGT
58	A_23_P16915	CATATTCCATTTTTAAGAAGAGGTGTTCCAGTTCTGCATCTGATACCGTCTCCTTTCCCT
60	A_23_P37076	GAAAATCCCTTGCTATGTCTTTCCTACTAGAAATGTTCTAGAATCGCTGGACGGTGGGGT
61	A_23_P166408	TGGTGGTGGATCCTGGAATTTTCTCACGCAGGAGCCATTGCTCTCCTAGAGGGGGTCTCA
63	A_23_P94501	GGCTCTTTGTGGAGGAAACTAAACATTCCCTTGATGGTCTCAAGCTATGATCAGAAGACT
67	A_23_P94230	TGAAGCTATTTCTGGGAGCCCAGAAGAAATGCTCTTTTGCTTGGAGTTTGTCATCCTACA
68	A_23_P154037	TTGGTTTCCTCTAGGGTGATATTCGTCATTACTCTGTCTCTTCAATCCATCCAGCTAAAT
69	A_23_P47938	AAGAAAACACACCTCGGCGACAATGTCTTGCTGCTCGGATTAGGTGGGGGATGGGCGACA
70	A_23_P90743	CTGAGTTTGCCTTGTTAATCTTCAATAGTTTTACCTACCCCAGTCTTTGGAACCCTAAAT
71	A_23_P171249	AAAGTGTCAAGTGATTAAGTGTGTATTTGTACCCTAGATGATATGAACCAGCAGTCTTGT
72	A_23_P142675	TGAGCTGTTCCCTTCTCTAAGCCATAATCTCTTAGTGGATTGAGCCCTCTTGGAAAGACT
73	A_23_P153637	TGTTATTGGCCTAGAGCTACACGTATATGGGTTTGTCCTGAGTCCGTTTTCAAATGACCT
76	A_23_P76749	CCTGTTCTGTTTTTGCTTTTCCTCTTCTTGACCAAAGCATGTGCCACTAGCTGTCCTTGA
80	A_23_P169061	TTGAAGGCAAAGATCATCAATATCTGCATCTGGCTGCTGTCGTCATCTGTTGGCATCTCT
91	A_23_P206253	CGAATTGGGAGGCTTATATTTTTCAGCAAAGAAATTTTGGGGGGTTTTGTGTTGTTGGGC
95	A_23_P151995	AATAAACAACTTTGATGATGTAACTTGACCTTCCAGAGTTATGGAAATTTTGTCCCCATG
97	A_23_P138011	AGAGACCTGCAGGGGCCTCGGCCCCTCACATCGTGTATGTCTCTCCTTGATTTGTGTTGT
100	A_23_P35912	GCCAAAGCTCAAATGCCCACCATAGAACGACTGTCCATGACAAGATATTTCTACCTCTTT
101	A_23_P99424	TGGCTCCCCATCATGTATCCTCCCGATTATTGCGTATTCTAAAATAGGAAACAAGACTTT
105	A_23_P418986	GATGACACTGCCACCTCTGACTTCTGCCTCTGGCCTTCCACTCTCAGTAAGAAGAGCCAG

Tabla 4: Sondas informativas para cáncer de mama – genes que no son de la familia (i) ni (ii)

Nº de sonda	Id. Agilent	
1	A_23_P366812	TTTACTTCTACCTGCTCTTCCCCAACTCCCTGAGCCTGAGTGAGCGTGTGGCCATCATCA
4	A_23_P389391	TGGGCCTCAAAATGGAGATGGATCCCAGGTCTTGTGGGACCCTGGGATGTTTGGGGACTT
7	A_23_P4096	TAATATCCCCAAACCTGAGATGAGCACTACGATGGCAGAGAGCAGCCTGTTGGACCTGCT
9	A_23_P15460	GACTGAAAAATCAGCTTTCTATTTACATGAAACACTTTGGGGGTCATGGGAGTGCACAGC
11	A_23_P379596	AGGGATAATTCAAACTGACAACCTGTGCAGTCCCGTGGAGGGTAGGGGAGTGTGGGTGAT
12	A_23_P391275	TAAATTATGATTTACTCTGTGCTGTTTCCAAATTGGGACCAGGAGAGAAATATGAACTTC
14	A_23_P124661	TCTATTATTTATAACTTCAGACTTGGGCCCCCTGTTCTTTCTTTCCCATTAACTTGAGTG
15	A_23_P44257	AACATTTTACTTCTGCGCTTCTATGTTTGGGAAACATTGCTCTGATAAAAAATAGCTGTC
21	A_23_P128183	CTGAGAGTTTTTGCAGAAATGGGGCAGAGGGACACCCTTTGGGCGTGGCTTCCTGGTGAT
22	A_23_P331211	CGAGTGGCTCACTCAGAATTCTTCATTGATGGGCTAGGGACCCTACTCGTGGGGTCATGC
23	A_23_P94932	TCTGTTGATGACCTTGGATGCTGTAAAGTGATTCGTCATAGTCTCTGGGGTACCCATGTA
25	A_23_P102122	AAGCGGCTGGCAACTGAAGGCTGGAACACTTGCTACTGGATAATCGTAGCTTTTAATGTT
30	A_23_P407654	GAGGAGCTCTTTTCTAGAGAGCCGGGAGTTGGGGAGGGGGTATTTATTTTGTTATTTATT
33	A_23_P392457	CCTCTGACTGCCTCCAACGTAAAAATGTAAATATAAATTTGGTTGAGATCTGGAGGGGGG
36	A_23_P406376	GCCACACTGGCTTTAGGACCTGTTGACACGGAGGGGGGTTTTTAATTTGGTTTTTAACAA
38	A_23_P22723	AACAAACTACAGTTTTACCGTGTGTTTGCCATTTGAGCTGTGTGGTGGGCAGGGGGCTGG
39	A_23_P70258	AGAGAGGATGGCTGTATCCTATCCCAGCTCAAGCTGCCAGCAGCAATGTTGGCTGCCCA
43	A_23_P104005	AATTTTCAAGACTTCTTTTCACTCTTTGATTTGGATCTGGCAAATTGGGGAGGGGATGCT
44	A_23_P119652	TTGCCCAACTGACCGTGGGCTGAACACACGTTCTGCTTGACTCATTTAGGGGGGAGGGAA
46	A_23_P22957	ATGAGGTGATCACTGTGTTCAGTGTTGTTGGAATGGATTCAGACTGGCTAATGGGGGAAA
48	A_23_P8072	GGGGGAGATCAGAATCGTCCAGCTGGGCTTCGACTTGGATGCCCATGGAATTATCTTCAC
50	A_23_P23346	AATCTTCTGAACGGCATAAGTCCTATTTTAGCCTTACCTCCTGCATTTGCAATACGTAAT
51	A_23_P92342	CGAACAAACAAAATACTTGGCGGGGCCCGAGAGGGCTCGTTTGGCCTATTCGTTGGGGAT
54	A_23_P153183	ACAGAAAACAGACTTGTAAAAAGCTTAGATCATCAAGTGTTTTGGATTGGGGGCCTCCCA
56	A_23_P157231	TGCAGAATGCATAAGATGAACATTGCATGACCGGATCATTTTAGTGTCTTTGCGTTAAAA
57	A_23_P103282	TGAAGATCATGAAGAAGCAGGGCCTCTACCTACAAAAGTGAATCTTGCTCATTCTGAAAT
59	A_23_P109462	CTGGATGTTTACCTGGAGACCGAGAGCCATGACGACAGTGTGGAGGGGCCCAAGGAATTT

ES 2 358 910 T3

62	A_23_P395460	TTAATGCTTTATACTGCCGAGTCTGGGGGCTTGTTTTGGTTTGGGGGCAGCCATCCTCCA
64	A_23_P418485	TCTAGGACTAATTCACACTGCAACAAAGGGGCTGATTAGAGCTTTTGAAGATGGGGGGAT
65	A_23_P215111	GACTTAACCACGTCAGAGGAAGGACTTTGGCAAGTGATATTGTCTTCATGTGGGGTATTA
66	A_23_P19543	CTGTCAAATTGCCACGATCTCACTAAAGGATTTCTATTTGCTGTCAGTTAAAAATAAAGC
74	A_23_P18317	TCATCTGCACTCAACATTTAATCGTGTCCTTGCTGTCTTTTTATTTTCCTTTTTGTTTGT
75	A_23_P89369	GCGGGAGGAGCGGCCGCTGATGGTGTTCAACGTCAAGTAGCGCCCGCGCAGGGCGGGGCA
77	A_23_P330561	CTGTCTCCCTGTTTGTGTAAACATACTAGAGTATACTGCGGCGTGTTTTCTGTCTACCCA
78	A_23_P206103	GAGAGTTTCTTTTAAATAATCAGCGGGTGTTGGTGATTTGTAGCCCTTCTGCCCTTAAAT
79	A_23_P98042	ATACTTTGTGAGTTCACCTGTCTTTATACTCAAAAGTGTCCCTTAATAGTGTCCTTGCCC
81	A_23_P166453	ACCTTTGAATTTGCGGATGCTGAGGAGGATGATGAGGTCAAGGTGTGAGGGGCTGGGGCA
82	A_23_P432554	TATTAGACTATGTCATCAATTTTTGCAAAGGTAAATTTGACTTCCTTGAACGGCTCTCAG
83	A_23_P368028	AAATACTGGGTGGCTTGGTTTAGAGCTAATTGTAGTGGAAGCCTGCAAGGTTGAGGGGTG
84	A_23_P213334	ACTCACTATGGCCAGAAAGCAATCTTGTTTCTCCCCCTGCCAGTCTCTTCTGATTAAAGA
85	A_23_P102113	AACAAATATTTATTTTGCACTCTCTTTGCGGCACTCTGGGGGCGGTGGGGTGCGTGGGGG
86	A_23_P319682	CAAGTTGTCACTGGAGATGCGCGCGGACTTGGCCCAAAACGTGCTTCTCTGCGGTGGGTC
87	A_23_P104471	GATTTCCCTGACCCAATTCAGAGATTCTTTATGCAAAAGTGAGTTCAGTCCATCTCTATA
88	A_23_P118749	GAAGGACTCGGTGATACCCACTGGGATCTTTTATCCTTTGTTGCAAAAGTGTGGACACTT
89	A_23_P420879	CAGGGCAACTCAAAGAATGTTCTGCTGGCATGTCCTATGAACATGTACCCGCATGGACGC
90	A_23_P29816	GAGAAAAGCAAAGCTCTTTCTTATTTTCCTCATAATCAGCTACCCTGGAGGGGAGGGAGA
92	A_23_P41992	TGAAATGCTGGAAGGGTTCTTCTCCCACAACCCCTGCCTCACGGAGGCCATTGCAGCTAA
93	A_23_P75479	AGACCTCGGTGATCACTGAGGGATTTCCGCGAGCTCGGCCTCACTTCTGCCCCGACTTGT
94	A_23_P307940	CTACAAGATTGGCAAAGAGATGCAGAATGCATAAGATGAACATTGCATGACCGGATCATT
96	A_23_P98910	AGGTTCTCAGAATGACCGTAAGATAGCTTACATTTCCTCTTTTTGCCTTTATCTCCCCAA
98	A_23_P149736	CCGTTTTGTTTCTGCTCAGTAATATAGTCAAGCAAGTTTGTTCCAAGTGACCCATTGAGC
99	A_23_P320250	AAATTGGCGCTGGAATTTGGGCTGGGAAAAATCTTGTGGTTATTTCCTTTAAAAAGGAAC
102	A_23_P55123	TCACGTTAACATATAGACACTGTTGGAAGCAGTTCCTTCTAAAAGGGTAGCCCTGGACTT
103	A_23_P251825	CTATGACACCTTTAAGGAGGTTGTTGGATCAGGGATGCAGTACCCACTTGCAGTCAAAAT
104	A_23_P109864	TGTGGGTGTCCAGCATCTTCTTCTTCCTTCCTGTCTTCTGTCTCACGGTCCTCTACAGTC
106	A_23_P428875	GTCGCCTGGGATTTTCATCCCTCGCACAAGGACTACGGGTTCACACGGTGAACTGGGGGA
107	A_23_P73466	GCCATAAGAAATTTGACAAGATGGTGGACACTCCTGCCTCCGAGCCTGCCCAAGCCTCCA
108	A_23_P106532	AAGGCCTTTGAGGTTGTGACTGTGGCTGGTATATCTGGCTGCCATTTTTCTGATGCATTT

109	A_23_P112251	AGAATTCTTAACTTCACAAGTGTTTTACTTCGACGATGTGCCTTTGATTTAATTTGGGAC
110	A_23_P313330	TCATTAGACATCGGGGATTTCACTCTGCAGAGTAATCCTGGAACTACATTAAAGTGGGGG
111	A_23_P27414	TGCGGGAAGCCTTTCAGCCACCGTTGCAACCTCAACGAGCACCAGAAGCGGCACGGGGGC
112	A_23_P210981	TTGTAGGACTTAATGGCTAAGAATTAGAACATAGCAAGGGGGCTCCTCTGTTGGAGTAAT

Tabla 5: Genes informativos para cáncer de mama – genes de las familias (i) y (ii)

Nº de sonda	Nº 1 de acceso	Nº 2 de acceso	Similitud génica y supuesta función biológica
			Factores de transcripción [familia (i)]
2	NM_006942	AB006867	Caja 20 de SRY (región Y determinante del sexo), un miembro de la familia de factores de transcripción que contiene la caja HMG relacionada con SRY
3	NM_002095	X63469	Factor IIE 2 de transcripción general (subunidad de 34 kDa), subunidad beta del factor de transcripción TFIIE de la RNA polimerasa II, necesario para el inicio de la transcripción; interacciona con activadores y con proteínas reparadoras de DNA; puede desempeñar un papel en la reparación acoplada a transcripción
6	NM_018942	M99587	Homeocaja 1 de H6, un miembro de la familia de proteínas ligantes de DNA que contienen homeodominios, un represor transcripcional que puede antagonizar la activación transcripcional mediada por Nkx2-5 de ratón
26		BC026031.1	Caja T 6, miembro de la familia de factores de transcripción con dominio caja T ligante de DNA; puede estar implicado en la somitogénesis y la formación del mesodermo paraxial embrionario
45	NM_005586	U78313	Inhibidor de la familia MyoD, un supuesto represor transcripcional que regula negativamente la miogénesis
69	NM_014212	AJ000041	Homeocaja C11, un factor de transcripción que contiene homeodominio; puede activar la transcripción dependiente de HNF-1 alfa (TCF1), puede actuar en el desarrollo y la diferenciación precoces del intestino; la proteína de fusión NUP98-HOXC11 está implicada en malignidades mieloides
105	NM_004348		Factor de transcripción 2 relacionado con Runt (RUNX2), de *Homo sapiens*, mRNA
			Genes relacionados con la defensa [familia (ii)]
63	NM_000700	BC035993	Anexina I, una proteína ligante de fosfolípidos dependiente de calcio, que inhibe la fosfolipasa A2 y ejerce actividad antiinflamatoria; implicada en la respuesta al estrés; asociada con el inicio precoz de tumorigénesis en carcinomas de esófago y próstata
8	NM_005139	M63310	Anexina A3 (lipocortina III), un miembro de la familia anexínica de proteínas ligantes de fosfolípido dependientes de calcio; se une a colina, ayuda a regular la permeabilidad y fusión de la membrana, y fagocitosis
13		BC007022.1	Amiloide sérico A1, una apolipoproteína de fase aguda que actúa en la quimiotaxis de leucocitos e induce metaloproteasas matriciales; puede desempeñar un papel en la artritis reumatoide, la aterosclerosis, la amiloidosis AA sistémica reactiva, la enfermedad de Alzheimer y la esclerosis múltiple
18	NM_004688	BC001266	Agente interactivo de N-myc (y STAT); proteína que interacciona con proteínas N-myc (MYCN) y STAT; aumenta la transcripción sensible a IL-2 e IFN gamma al promover la asociación de CBP/p300 con proteínas STAT; puede ayudar a BCRA1 a suprimir la carcinogénesis de cáncer de mama
19	NM_203503	AF325460	Miembro 11 de la superfamilia de lectinas de tipo C (dominio de reconocimiento de hidratos de carbono, dependiente de calcio), una glicoproteína de células dendríticas que inhibe la inducción por interferones alfa y beta y puede mediar en la captura antigénica para el inicio de respuestas inmunes dependientes de células T
27	NM_020387	AF274025	Proteína con alta similitud con la pequeña proteína 11a ligante de GTP, de tipo Ras p21 (RAB11A humana), que es una supuesta GTPasa que está implicada en la fagocitosis y posiblemente en el transporte de vesículas; miembro de la superfamilia Ras de proteínas ligantes de GTP
29	NM_012218	AJ271747	Factor 3 ligante de potenciadores de interleucinas, una subunidad de NF-AT; actúa como regulador transcripcional positivo o negativo, necesario para la expresión de IL-2 por células T, posiblemente implicado en el procesamiento de mRNA, la inhibición de la traducción, la defensa del huésped y la autoinmunidad
32	NM_181640	BC004380	Factor 1 de tipo quimiocina, un agente quimioatrayente secretado para leucocitos, neutrófilos, monocitos y linfocitos; estimula la res-

			puesta inflamatoria y la proliferación de células madre musculares; desempeña un papel en la regulación de la miogénesis
37	NM_153633	X07495	Homeocaja C4, un miembro de la familia de genes homeocaja de proteínas ligantes de DNA; puede desempeñar un papel en la regulación de la activación de linfocitos y la determinación del linaje durante la hematopoyesis
61	NM_020530	BC011589	Oncostatina M, un miembro de la familia citocínica de la interleucina 6, producida por linfocitos T y monocitos activados; regula el crecimiento y la diferenciación celulares a través de la activación de las rutas de JAK-STAT y MAPK; regula el crecimiento de células de sarcoma de Kaposi
67	NM_015364	AB018549	Proteína MD-2, parte de un complejo receptor del polisacárido; contribuye a la señalización del polisacárido y el receptor 4 de tipo Toll (Tlr4); puede desempeñar un papel en la respuesta de defensa celular
80	NM_000912	L37362	Receptor kappa 1 de opioides, un receptor acoplado a la proteína G que genera señales a través de una proteína G inhibitoria; puede modular una percepción sensorial tal como el dolor; su expresión alterada se asocia con la enfermedad de Alzheimer; la estimulación por agonistas puede suprimir infecciones por el VIH.
95	NM_004049	U29680	Proteína A1 relacionada con BCL2, un miembro de la familia Bcl-2 de reguladores de la apoptosis; inhibe la apoptosis, promueve la tumorigénesis y puede desempeñar un papel protector durante la inflamación
17	NM_003580	BC041124	Factor asociado con la activación de la esfingomielinasa neutra (N-Smasa); media en la inducción de la N-Smasa por el receptor CD40 del factor de necrosis tumoral; implicada en la inducción de apoptosis mediada por el TNF alfa; se une al receptor TNF-R55 del TNF (TNFRSF1A)
34	NM_144615	BC015655	Proteína que contiene un dominio de inmunoglobulina (Ig), que puede estar implicada en interacciones proteína-proteína y proteína-ligando
			Remodelación de la cromatina [familia (ii)]
10	NM_003529	BC067491	Miembro A de la familia H3 de histonas, un componente de los nucleosomas, junto con las histonas nucleares H2A, H2B y H4 y DNA
20	NM_003536	BC062305	Histona 1 de *Homo sapiens*, H3h (HIST1H3H), mRNA
31	NM_018282	AK090873	Proteína 1 de paraspeckle, una supuesta proteína ligante de RNA que contiene dos dominios ligantes de RNA (RRM); se mueve entre el compartimento de paraspeckles del espacio intercromatínico y el nucleolo e interacciona con el nucleolo de un modo dependiente de la transcripción
			Biogénesis ribosómica [familia (i)]
16		AK024156	Proteína de función desconocida, presenta una similitud moderada con una región de Bms1p de *S. cerevisiae*, que está implicada en el procesamiento de rRNA y la biogénesis de subunidades ribosómicas 40S
			Metabolismo proteico [familia (i)]
52	NM_139026	AY358118	Una proteína de tipo desintegrina y metaloproteasa (tipo reprolisina) de *Homo sapiens* con un motivo de tipo 1 de trombospondina, 13 (ADAMTS13), variante 1 de transcrito, mRNA
72	NM_012100	AK001777	Aspartil aminopeptidasa citosólica, un miembro de la familia M18 de metaloproteasas; tiene una preferencia de sustrato hacia restos de aspartilo y glutamilo N-terminales y puede estar implicada en el metabolismo peptídico intracelular
100	NM_001225	U28979	Caspasa 4, un miembro de la familia de cisteína proteasas ICE que está implicado en la inducción de apoptosis; la inhibición de la serpina CrmA del virus de la viruela bovina puede facilitar la infección al inhibir la apoptosis
101	NM_003291	AK097678	Tripeptidil peptidasa II, una serina exopeptidasa que puede actuar en el recambio proteico no proteasomal; neuropéptidos y antígenos del MHC de clase I son sustratos; actúa en la apoptosis activada por *Shigella*; suprarregulada en células de linfoma de Burkitt que sobreexpresan MYC
5	NM_012265	BC002705	Miembro de la familia romboidal de proteínas integrales de membrana; contiene un dominio UBA (asociado a ubiquitina) o TS-N
55		M64247.1	Proteína con similitud muy acusada con la troponina cardiaca I (Tnni3 de ratón), que es la subunidad inhibitoria de la troponina,

			miembro de la familia de troponinas, que regula la contracción muscular inducida por calcio
76		BC036812	Proteína con elevada similitud con la UDP-N-acetil-alfa-D-galactosamina:polipéptido N-acetilgalactosaminiltransferasa (GALNT2 humana), miembro de la familia 2 de glicosil transferasas; contiene 2 dominios QxW (ricina B) de repetición de lectina
			Estrés oxidativo [familia (ii)]
58	NM_012413	X71125	Glutaminil-péptido ciclotransferasa (glutaminil ciclasa), expresada en la hipófisis; la expresión en el epitelio del cristalino está infrarregulada durante el estrés oxidativo
68	NM_001159	L11005	Aldehído oxidasa, una flavoenzima que contiene molibdeno implicada en el metabolismo de radicales oxigenados, sustancias xenobióticas y fármacos; candidato génico para la causa de la forma recesiva autosómica de la esclerosis lateral amiotrófica
60	NM_004873	AK023145	Atanogén asociado con BCL2, contiene un dominio BAG; se pronostica que regula proteínas Hsc70/Hsp70 al unirse a sus dominios ATPasa a través de su dominio BAG
			Secreción de proteínas [síntesis de proteínas – familia (i)]
24	NM_012430	AF100749	Homólogo de Sec22, un miembro de la familia SEC22 de proteínas implicadas en el tráfico de vesículas; puede estar implicado en el tráfico de proteínas desde el retículo endoplásmico hasta el aparato de Golgi
28		M65199	Endotelina 2, un miembro de una familia de hormonas peptídicas vasoactivas, implicada en la regulación de la presión sanguínea; inhibe la secreción de prolactina, puede actuar en el crecimiento celular relacionado con el desarrollo del corazón; la multiplicación del locus se correlaciona con la hipertensión
35	NM_001661	L38490	Tipo factor 4 de ribosilación de ADP, una GTPasa y miembro de la familia de factores de ribosilación de ADP; puede estar implicado en el transporte vesicular intracelular y la secreción de proteínas
41		BC028121	Proteína con elevada similitud con la proteína de membrana que se asocia con la cadena de translocación (TRAM humana), que es un supuesto receptor del retículo endoplásmico que estimula la translocación de proteínas secretoras, miembro de la familia de proteínas de garantía de longevidad (LAG1)
42	NM_030772	AF271261	Miembro de la familia conexínica de proteínas de canales de unión en hendidura, que permiten el paso intercelular de moléculas; tiene una similitud moderada con la proteína alfa 1 de unión en hendidura (conexina 43, GJA1 humana), que se asocia con la heterotaxia visceroatrial
49	NM_013248	AK026360	Proteína exportadora 1 de tipo NTF2; se une a RAN; actúa en la ruta de exportación nuclear dependiente de CRM1 (XPO1)
53	NM_012346		Glicoproteína p62 de poro nuclear, un componente del poro nuclear; puede estar implicada en el transporte nucleocitoplásmico; diana para degradación durante una infección por poliovirus
73	NM_032139	AL136784	Proteína que contiene un dominio de la proteína 9 de selección vacuolar (VPS9) y ocho repeticiones de ankirina (Ank); tiene una región de baja similitud con una región de UNC-44 de *C. elegans*, que se requiere para la guía axonal y la apropiada fasciculación de axones
91		AB010419.1	Subunidad 2 del dominio runt alfa del factor ligante del núcleo, translocada a 3, miembro de la familia de la proteína MTG8 (ETO/CDR), supuesto factor de transcripción; en la leucemia mieloide aguda se ve una fusión del gen correspondiente con RUNX1
97		AL137537	Proteína con elevada similitud con la ATPasa transportadora de aminofosfolípidos (colestasis intrahepática familiar 1, ATP8B1 humana), que se asocia con la colestasis intrahepática familiar, miembro de la familia de la haloácido deshalogenasa o la epóxido hidrolasa
			Morfogénesis de células B [respuesta inmune, familia (ii)]
70	NM_002909	M27190	1 alfa derivada de islotes en regeneración (proteína del cálculo pancreático); induce la regeneración de células pancreáticas beta, mejora la diabetes en animales; su expresión aberrante se asocia con la pancreatitis calcificante crónica y la carcinogénesis de colon
71	NM_001551	BC004137	Proteína 1 ligante de inmunoglobulinas; puede estar implicada en la transducción de señales de células B mediada por el receptor de IgG

40		AC007032	Factor potenciador de colonias de células pre-B, una citocina que produce sinergia en la actividad de formación de colonias del factor de células madre (KITLG) y la interleucina 7 (IL-7) en células del linaje B precoces; puede desempeñar un papel en el parto prematuro inducido por infección y en el cáncer colorrectal primario
			Respuesta inmune [familia (ii)]
47	NM_152547	AK057097	Proteína con baja similitud con el homólogo 3 de B7 (B7-H3 humano), que es una molécula coestimulante para células T que regula positivamente la síntesis de interferón gamma y la proliferación y es inducida por citocinas inflamatorias

Los números de acceso 1 y 2 proporcionan números de acceso alternativos para el gen. La secuencia relevante puede ser identificada en la base de datos del NCBI (www.ncbi.nim.nih.gov).

Tabla 6: Genes informativos para cáncer de mama – genes que no son de la familia (i) ni (ii)

Nº de sonda	Nº 1 de acceso	Nº 2 de acceso	Similitud génica y supuesta función biológica
			Canales y bombas
1	NM_001651	BC034356	Aquaporina 5 de *Homo sapiens* (AQP5), mRNA
4	NM_005072	AF054506	Familia 12 de vehículos de solutos de *Homo sapiens* (transportadores de potasio/cloruro), miembro 4 (SLC12A4), mRNA
11	NM_004983	U52152	Canal rectificador interno de potasio de *Homo sapiens*, subfamilia J, miembro 9 (KCNJ9), mRNA
38		BC047580	ATPasa 3 de transporte de Ca^{2+} de la membrana plasmática; se ha pronosticado que está implicada en el transporte de calcio; expresada predominantemente en el cerebro
107	NM_174873	AF260427	Receptor purinérgico P2X2, un canal catiónico activado por ATP extracelular que está implicado en el transporte de iones calcio y la transducción de señales
65	NM_130840	AK055789	Isoforma 4 de la subunidad A de la ATPasa lisosómica V0 (que transporta H^+), subunidad 1B accesoria no catalítica de la bomba vacuolar de protones
44		AC004659	Transportador 4 de aminoácidos excitantes (miembro 6 de la familia 1 de vehículos de solutos), un transportador de glutamato y aspartato de alta afinidad con actividad de canal de cloruro activado por ligandos; regula probablemente la neurotransmisión excitante dentro del cerebelo
74	NM_017836	AK000480	Miembro de la familia de transportadores de cationes divalentes, que puede transportar Mg^{2+} u otros cationes divalentes en la célula; tiene una gran similitud con DKFZP434K0427 humano no caracterizado
			Supuestas proteínas cinasas o proteínas que interaccionan con cinasas
48	NM_032454	L26260	Serina treonina cinasa 19, una proteína cinasa dependiente de manganeso que se localiza principalmente en el núcleo
54		AK056549	Proteína 1 asociada a la membrana que interacciona con guanilato cinasa, proteína con acusada similitud con Maguin1 de rata, que contiene dominios de SAM, PDZ y PH e interacciona con las cinasas de andamiaje sináptico S-SCAM y PSD-95/SAP90
66	NM_003137	BC038292	Proteína cinasa para la familia de factores de corte y empalme de RNA rica en serinas y argininas (SR); actúa probablemente para controlar la localización de los factores de corte y empalme dentro del núcleo; puede desempeñar un papel en la determinación de la sensibilidad a cisplatina, un agente anticanceroso muy usado
78	NM_015518	BC056423	Proteína que contiene un dominio de proteína cinasa; tiene una similitud moderada con una región de la cinasa 1 de tipo unc-51 (Ulk1 de ratón), que es una proteína cinasa implicada en las fases tempranas de la extensión de neuritas de células granulares cerebelosas y puede actuar en cascadas de señalización
81	NM_007061	BC009356	Proteína medular-estromal-endotelial componente del suero; contiene un dominio CRIB no cinasa (interactivo con-ligante de

			Cdc42/Rac), se une a CDC42 de un modo dependiente de GTP, actúa en la reorganización del citoesqueleto y posiblemente en la transducción de señales de la proteína Rac
109	NM_004125	BC016319	Subunidad 10 de la proteína gamma ligante de nucleótidos de guanina, supuesto componente de complejos heterotrímeros de proteína G que están implicados en la transducción de señales, interacciona con las proteínas G beta 1 (GNB1) y beta 2 (GNB2) y con la cinasa supresora murina de Ras
			Metabolismo
7	NM_000717	M83670	Anhidrasa carbónica IV, cataliza la hidratación reversible del dióxido de carbono para formar bicarbonato y un protón, desempeña un papel en la regulación del pH, puede actuar en la absorción renal de bicarbonato; su deficiencia se puede asociar con la acidosis tubular renal proximal pura
22	NM_153446	AJ517771	1,4-N-acetilgalactosaminiltransferasa beta (GALGT2) de *Homo sapiens*, mRNA
102	NM_001303	U09466	Hemo A:farnesiltransferasa, una farnesiltransferasa necesaria para la biosíntesis de hemo A; la deficiencia o alteración del gen se puede asociar con la neuropatía hereditaria con tendencia a parálisis por presión y enfermedad Charcot-Marie-Tooth de tipo 1
108	NM_130468	BC023653	Dermatano-4-sulfotransferasa-1, cataliza la transferencia de un sulfato al hidroxilo C-4 de la N-acetilgalactosamina de dermatano en la biosíntesis de sulfato de dermatano
			Relacionados con el cáncer
21	NM_145897	D89667	Prefoldina 5, un componente del complejo de la chaperona prefoldina implicado en la distribución de proteínas no plegadas a la chaperonina citosólica; interacciona con MYC y puede reprimir su activación; candidato a supresor tumoral comúnmente sustituido en células cancerosas
56	NM_006136	BC005338	Proteína de remate (capping) de la línea Z (alfa 2), subunidad de una proteína ligante de actina que puede desempeñar un papel en la motilidad celular; el gen correspondiente está multiplicado en gliomas malignos y puede estar implicado en tumorigénesis
79	NM_004728.1		Polipéptido 21 de la caja DEAD-H (Asp-Glu-Ala-Asp/His), una RNA helicasa que resulta inhibida por el fármaco anticanceroso adriamicina, una RNA foldasa que introduce una estructura secundaria intramolecular en el ssRNA, un autoantígeno en la enfermedad del estómago en sandía
83		AF010315	Proteína 11 de *Homo sapiens* inducible por la proteína tumoral p53 (TP53I11), mRNA
84	NM_033137	X65779	Factor 1 (ácido) de crecimiento de fibroblastos, un mitógeno e inhibidor de apoptosis implicado en la migración celular, la embriogénesis, el desarrollo de órganos y la angiogénesis
85	NM_025216	AK024363	Miembro 10a de la familia de sitios de integración de MMTV, de tipo sin alas; miembro de la familia wnt, puede estar implicado en la transducción de señales y la carcinogénesis; sobreproducido en ciertos cánceres esofágicos, gástricos y colorrectales
93	NM_021070	AF318354	Proteína que contiene ocho dominios de tipo factor de crecimiento epidérmico (EGF) y dos dominios de proteína ligante de TGF; tiene una acusada similitud con una región de la proteína 3 ligante del factor de crecimiento transformante latente (Ltbp3 de ratón)
103	NM_001550	BC001272	Proteína con una acusada similitud con Rn.3723 de rata, que es inducida por el factor de crecimiento nervioso (NGF); desempeña un papel en la diferenciación muscular y se expresa en tejidos en proliferación y diferenciación
104	NM_198407	U60179	Receptor secretagogo de la hormona de crecimiento, un receptor acoplado a la proteína G que se une a grelina (GHRL) y secretagogos de la hormona de crecimiento sintéticos; puede regular la secreción de hormona de crecimiento; su expresión elevada se puede asociar con tumores endocrinos
88	NM_018041	BC010890	F-LAN-1, proteína suprarregulada en hepatocarcinomas; implicada en la regulación positiva de la proliferación celular
			Relacionados con actina
25	NM_005731	U50523	Subunidad 2 del complejo de la proteína 2/3 relacionada con actina, componente del complejo Arp2/3, que está implicada en la ensambladura del citoesqueleto de actina; interacciona directamente con ARPC4, posiblemente como un producto intermedio precoz, en la formación del complejo Arp2/3

98	NM_006135	BX648738	Proteína alfa 1 de remate de la línea muscular Z, una proteína rematadora de actina que regula la polimerización de actina y puede contribuir al remate de la actina en forma de cola de flecha, la motilidad celular, la organización de sarcómeros y la función muscular
94	NM_006136	U03269	Proteína de remate (filamento de actina) de la línea muscular Z de *Homo sapiens*, alfa 2 (CAPZA2), mRNA
			Diferenciación celular
39	NM_001858	U09279	Subunidad alfa 1 del colágeno de tipo XIX, miembro de la familia de colágenos FACIT; puede estar implicado en la diferenciación celular; alternativamente cortado y empalmado en células de rabdomiosarcoma
50	NM_006818	AK056089	Gen del cromosoma 1q, fusionado con ALL1; proteína expresada en el timo y líneas celulares hematopoyéticas y leucémicas; el gen correspondiente es el sitio de translocaciones cromosómicas que afectan a MLL y dan lugar a la leucemia mielomonocítica aguda
51	NM_006168		Familia 6 A de genes homeocaja NK, un miembro de la familia de homeodominios de proteínas ligantes de DNA que regulan la expresión génica y participan en el control de la diferenciación celular; contiene regiones muy conservadas de homeodominios y del decapéptido NK
92	NM_001496	AY359037	Receptor alfa 3 de la familia del GDNF, un miembro huérfano de la familia de receptores de GDNF/neurturina/persefina, ligado a glicosilfosfatidilinositol (GPI); muy expresado en el sistema nervioso periférico en desarrollo y en ganglios sensoriales y simpáticos adultos
			Otras funciones
46	NM_016009	AK001954	Endofilina B1 de tipo GRB2 con dominio SH3; contiene un dominio de homología Src 3 (SH3) en el extremo C y puede actuar como regulador de la ruta de señalización apoptótica de BAX
89	NM_021724	M24898	Subfamilia 1 de receptores nucleares de *Homo sapiens*, grupo D, miembro 1 (NR1D1), mRNA
96	NM_006152	U10485	Proteína de membrana linfoidemente restringida, una proteína de membrana de la cara citoplásmica del retículo endoplásmico
90	NM_016364	BC009778	Fosfolipasa A1 específica de fosfatidilserina; hidroliza ácidos grasos en la posición sn-1 de la fosfatidilserina y la 1-acil-2-lisofosfatidilserina; desempeña un papel en la regulación de funciones mediadas por fosfatidilserina o lisofosfatidilserina
87	NM_015900	BC047703	Fosfatasa 13 de especificidad doble; puede desfosforilar restos de fosfotirosina, fosfoserina y fosfotreonina; puede desempeñar un papel en la regulación de la meiosis y/o la diferenciación de células germinales testiculares
			Función desconocida
9	NM_018286	AK095175	Proteína de función desconocida
12	NM_013441	AF176117	2 de tipo gen 1 de la región crítica del síndrome de Down de *Homo sapiens* (DSCR1L2), mRNA
14	NM_148415		Proteína con una similitud moderada con SCA2 (ataxina 2)m que se asocia con la ataxia espeinocerebelosa de tipo 2
15	NM_017845	BC015145	Proteína de función desconocida, presenta una alta similitud con D5Buc26e no caracterizada de ratón
23	NM_015702	BC022859	Proteína de función desconocida, presenta una alta similitud con 2010311D03Rik no caracterizada de ratón
30	NM_173564	AK124773	Hipotética proteína FLJ37538 (FLJ37538) de *Homo sapiens*, mRNA
33	NM_002336	AK074543	Proteína 6 relacionada con el receptor de lipoproteínas de baja densidad de *Homo sapines* (LRP6), mRNA
36	NM_152383	BC036113	Hipotética proteína MGC42174 (MGC42174) de *Homo sapiens*, mRNA
43	NM_020141	AF164793	Proteína de función desconocida, presenta una alta similitud con K07F5.15 no caracterizada de *C. elegans*
57	NM_004872	BC016374	Proteína de función desconocida, presenta una acusada similitud con ORF18 no caracterizada de ratón
59	NM_003678	AK025385	Proteína de función desconocida, presenta una acusadísima similitud con Fmip no caracterizada de ratón
62	NM_004321	BX537556	Hipotética proteína BC009491 (LOC151568) de *Homo sapiens*, mRNA
64	NM_152587	BC029536	Hipotética proteína MGC33948 (MGC33948) de *Homo sapiens*, mRNA
75		BC030200.1	Proteína de función desconocida, presenta una baja similitud con D430039N05Rik no caracterizada de ratón
77	NM_174918	BC035847	Hipotética proteína LOC199675 (LOC199675) de *Homo sapiens*, mRNA

82	NM_174899	BC033935	Hipotética proteína LOC130888 (LOC130888) de *Homo sapiens*, mRNA
86	NM_178525	AY248901	Hipotética proteína MGC33407 (MGC33407) de *Homo sapiens*, mRNA
99	NM_025109	AL133017	Hipotética proteína FLJ22865 (FLJ22865) de *Homo sapiens*, mRNA
106	NM_152362	AK024161	Hipotética proteína MGC17791 (MGC17791) de *Homo sapiens*, mRNA
110		XM_088567	Desconocido
111	NM_198458	AK126727	Desconocido
112		BC054888	Desconocido

Los números de acceso son como se definieron en la Tabla 5.

REIVINDICACIONES

1. Un método para preparar un patrón estándar de transcritos génicos característico del cáncer de mama en un organismo, que comprende al menos las operaciones de:

a) aislar mRNA de las células de una muestra de uno o más organismos que tienen cáncer de mama;

b) hibridar el mRNA de la operación (a) con un conjunto de sondas oligonucleotídicas específicas para dicho cáncer de mama en un organismo y una muestra del mismo que corresponda al organismo y una muestra del mismo bajo investigación, en que dicho conjunto consiste en menos de 500 oligonucleótidos y comprende todos los oligonucleótidos enumerados en la Tabla 2 ó 3, en que cada oligonucleótido puede ser sustituido por un oligonucleótido que es una parte de dicho oligonucleótido de la Tabla 2 ó 3 o por un oligonucleótido con una secuencia complementaria; y

c) evaluar la cantidad de mRNA que se hibrida con cada una de dichas sondas para producir un patrón característico que refleje el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en la muestra con cáncer de mama.

2. Un método para preparar un patrón de transcritos génicos de ensayo, que comprende al menos las operaciones de:

a) aislar mRNA de las células de una muestra de dicho organismo de ensayo;

b) hibridar el mRNA de la operación (a) con un conjunto de oligonucleótidos como los definidos en la Reivindicación 1, específicos para el cáncer de mama en un organismo y una muestra del mismo que corresponda al organismo y una muestra del mismo bajo investigación; y

c) evaluar la cantidad de mRNA que se hibrida con cada una de dichas sondas para producir dicho patrón que refleja el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en dicha muestra de ensayo.

3. Un método para diagnosticar o identificar o controlar el cáncer de mama en un organismo, que comprende las operaciones de:

a) aislar mRNA de las células de una muestra de dicho organismo;

b) hibridar el mRNA de la operación (a) con un conjunto de oligonucleótidos como los definidos en la Reivindicación 1, específicos para cáncer de mama en un organismo y una muestra del mismo que corresponde al organismo y una muestra del mismo bajo investigación;

c) evaluar la cantidad de mRNA que se hibrida con cada una de dichas sondas para producir un patrón característico que refleje el nivel de expresión génica de los genes a los que se unen dichos oligonucleótidos, en dicha muestra; y

d) comparar dicho patrón con un patrón diagnóstico estándar preparado de acuerdo con el método de la Reivindicación 1, utilizando una muestra de un organismo que corresponde al organismo y una muestra bajo investigación para determinar la presencia de cáncer de mama en el organismo bajo investigación.

4. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 3, en que dicho mRNA aislado es inversamente transcrito a cDNA y dicho cDNA sustituye a dicho mRNA en las operaciones b) y c) de dicho método.

5. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 4, en que dicho conjunto comprende además todos los oligonucleótidos enumerados en la Tabla 4, en que cada oligonucleótido puede ser sustituido por un oligonucleótido que es una parte de dicho oligonucleótido de la Tabla 4 o por un oligonucleótido con una secuencia complementaria.

6. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 5, en que dicho conjunto de sondas está inmovilizado sobre uno o más soportes sólidos.

7. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 6, en que dichas células no son células de la enfermedad, no han entrado en contacto con células de la enfermedad y no proceden del sitio de la enfermedad.

8. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 7, en que dicha muestra ha sido obtenida de un sitio alejado del sitio de la enfermedad.

9. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 8, en que dicha muestra es tejido, fluido corporal o desecho corporal.

10. Un método como el reivindicado en la Reivindicación 9, en que dicha muestra es sangre periférica.

11. Un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 10, en que dicho organismo es un mamífero.

12. Un método como el reivindicado en la Reivindicación 11, en que dicho mamífero es un ser humano.

13. Un conjunto de menos de 500 sondas oligonucleotídicas como el definido en cualquiera de las Reivindicaciones 1 a 12.

14. Un kit para llevar a cabo un método como el reivindicado en cualquiera de las Reivindicaciones 1 a 12, que comprende un conjunto de sondas oligonucleotídicas como el definido en la Reivindicación 13, inmovilizado sobre uno o más soportes sólidos.

15. Un kit como el reivindicado en la Reivindicación 14, que comprende además una información añadida que detalla cómo se debería llevar el método a cabo.

16. El uso de un conjunto de sondas oligonucleotídicas o un kit como los definidos en cualquiera de las Reivindicaciones 13 a 15 para determinar el patrón de expresión génica de una célula, patrón que refleja el nivel de expresión génica de genes a los que se unen dichas sondas oligonucleotídicas para el diagnóstico del cáncer de mama, que comprende al menos las operaciones de

a) aislar mRNA de dicha célula;

b) hibridar el mRNA de la operación (a) con un conjunto de sondas oligonucleotídicas o un kit como los definidos en cualquiera de las Reivindicaciones 13 a 15; y

c) evaluar la cantidad de mRNA que se hibrida con cada una de dichas sondas para producir dicho patrón.

17. Un uso como el reivindicado en la Reivindicación 16, en que dicho mRNA aislado es inversamente transcrito a cDNA y dicho cDNA sustituye a dicho mRNA en las operaciones b) y c) de dicho método.

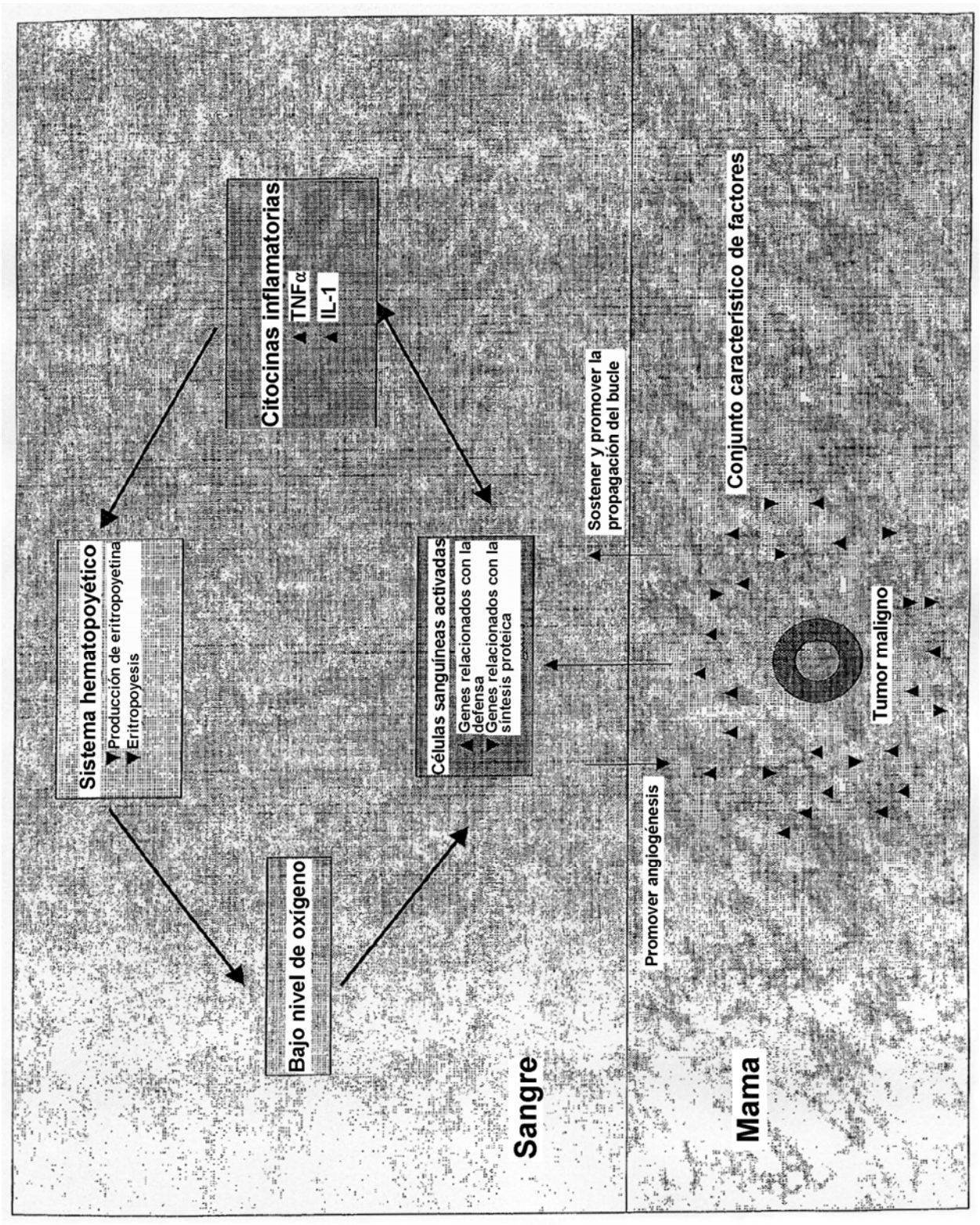
Sistema hematopoyético
Producción de eritropoyetina
Eritropoyesis
Citocinas inflamatorias
TNFα
IL-1
Bajo nivel de oxígeno
Células sanguíneas activadas
Genes relacionados con la defensa
Genes relacionados con la síntesis proteica
Sostener y promover la propagación del bucle
Conjunto característico de factores
Promover angiogénesis
Tumor maligno
Sangre
Mama

Figura 1

Gráfico de calificación basado en 35 genes usando PLSR

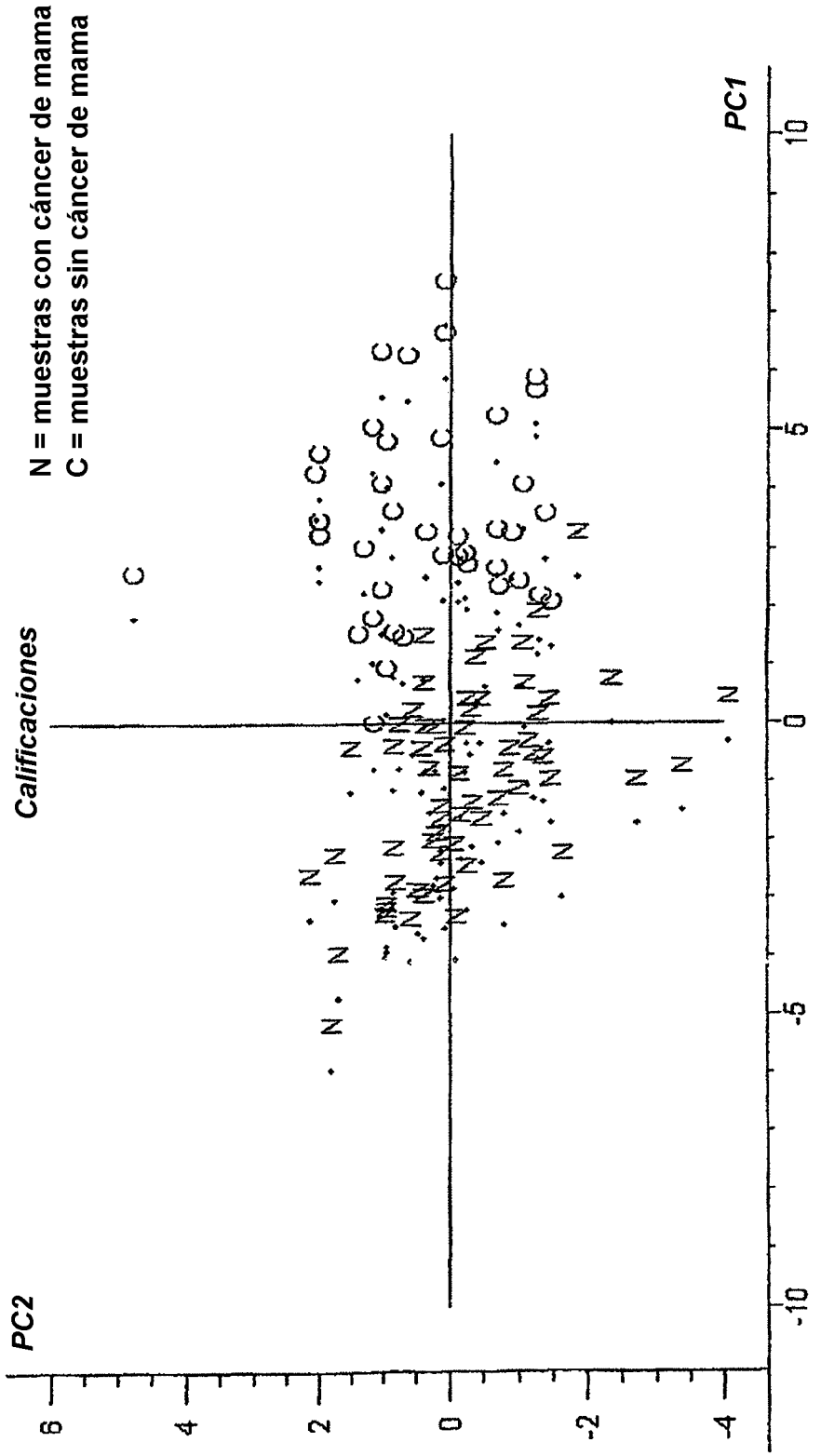

Figura 2

Gráfico de predicción basado en 35 genes usando PLSR

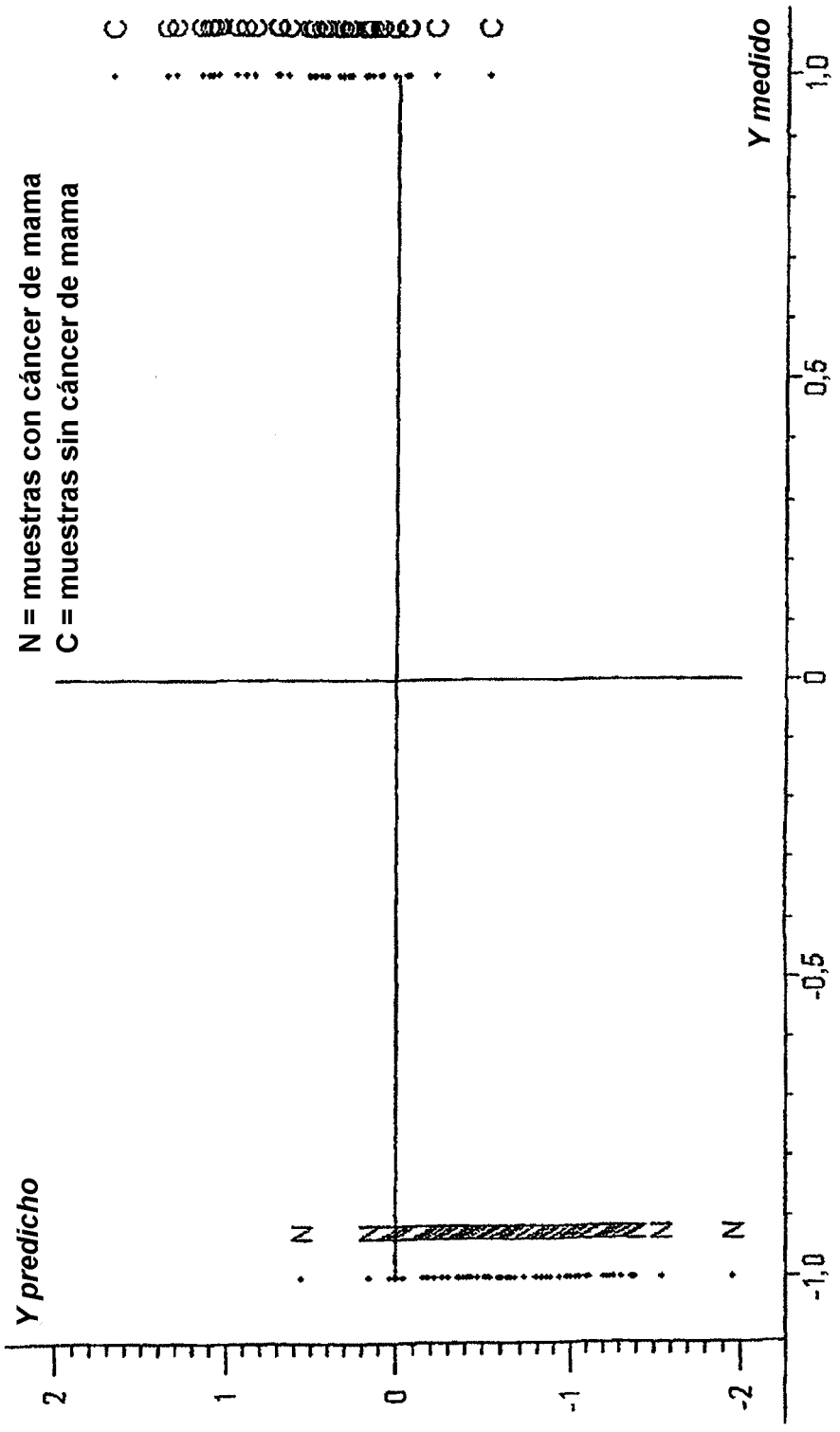

Figura 3

Nivel de expresión medio de 35 genes en muestras con cáncer de mama y sin cáncer de mama. Los genes se representan por sus Ids. de posición en la red.

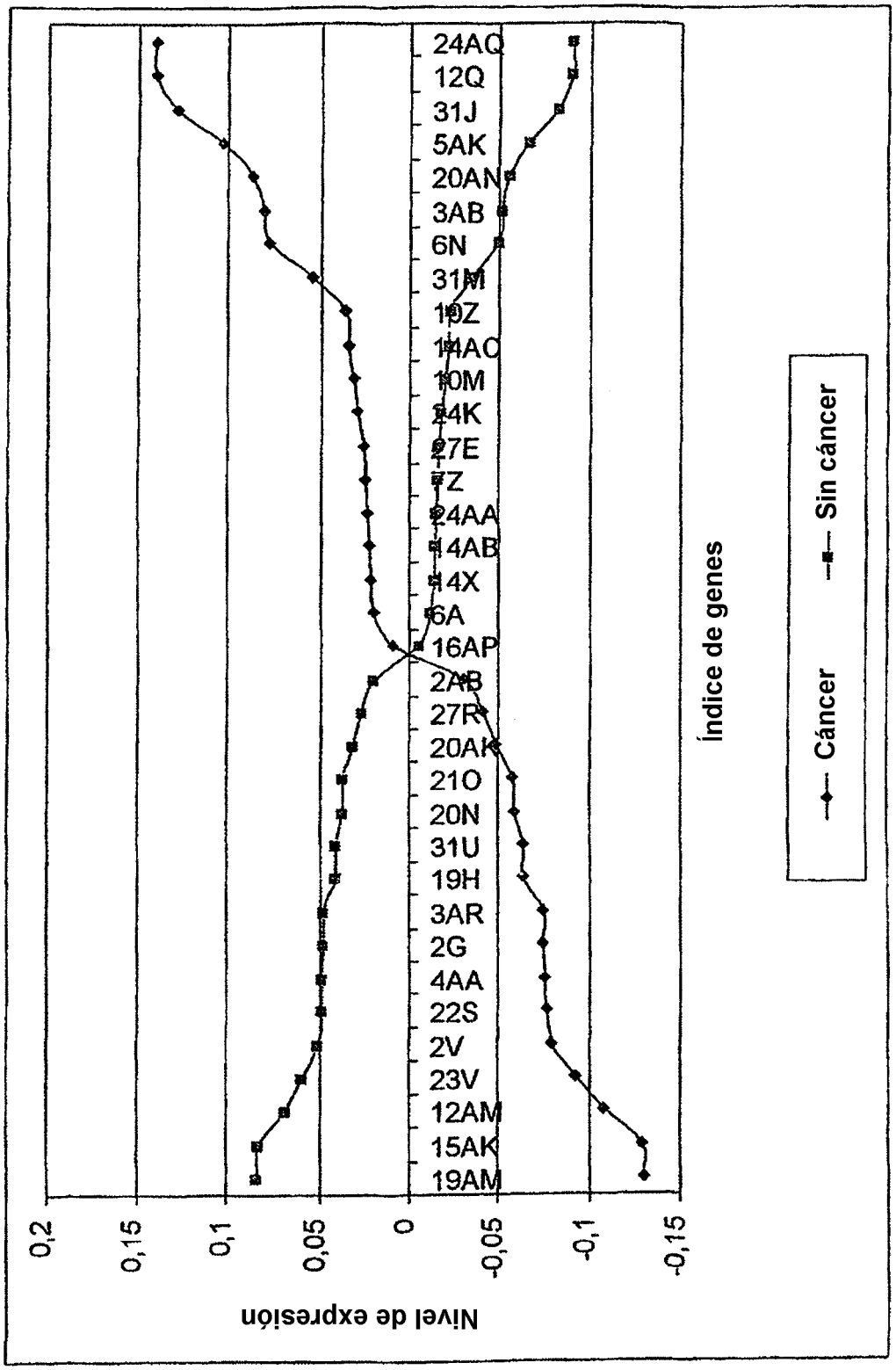

Figura 4