

19



OFICINA ESPAÑOLA DE  
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 700 617**

51 Int. Cl.:

**C12Q 1/68** (2008.01)

**G06F 19/24** (2011.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

96 Fecha de presentación y número de la solicitud europea: **19.07.2012 E 16192584 (7)**

97 Fecha y número de publicación de la concesión europea: **19.09.2018 EP 3173489**

54 Título: **Evaluación de la actividad de la ruta de señalización celular utilizando un modelo probabilístico de la expresión del gen objetivo**

30 Prioridad:

**19.07.2011 US 201161509137 P**  
**19.08.2011 EP 11178148**

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

**18.02.2019**

73 Titular/es:

**KONINKLIJKE PHILIPS N.V. (100.0%)**  
**High Tech Campus 5**  
**5656 AE Eindhoven, NL**

72 Inventor/es:

**VERHAEGH, WILHELMUS, FRANCISCUS,**  
**JOHANNES;**  
**VAN DE STOLPE, ANJA;**  
**VAN OOIJEN, HENDRIK JAN;**  
**DULLA, KALYANA, CHAKRAVARTHI;**  
**ALVES DE INDA, MARCIA y**  
**HOFFMANN, RALF**

74 Agente/Representante:

**ISERN JARA, Jorge**

**Observaciones:**

**Véase nota informativa (Remarks, Remarques o Bemerkungen) en el folleto original publicado por la Oficina Europea de Patentes**

**ES 2 700 617 T3**

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

## DESCRIPCIÓN

Evaluación de la actividad de la ruta de señalización celular utilizando un modelo probabilístico de la expresión del gen objetivo

5 El objeto descrito en el presente documento se refiere principalmente a bioinformática, a técnicas de procesamiento genómico, a técnicas de procesamiento proteómico y a técnicas relacionadas.

10 Los análisis genómicos y proteómicos tienen perspectivas sustanciales realizadas y potenciales para la aplicación clínica en campos médicos tales como la oncología, donde se sabe que diversos cánceres están asociados con combinaciones específicas de mutaciones/variaciones genómicas y/o niveles de expresión altos o bajos para genes específicos, que juegan un papel en el crecimiento y la evolución del cáncer, por ejemplo proliferación celular y metástasis. Por ejemplo, la ruta de señalización de Wnt afecta a la regulación de la proliferación celular, y está altamente regulada. La alta actividad de la ruta Wnt debida a la pérdida de la regulación se ha correlacionado con el cáncer, entre ellos con tumores malignos de colon. Sin limitación a ninguna teoría en particular de operación, se cree que la desregulación de la ruta Wnt en células de colon maligno conduce a la alta actividad de la ruta Wnt que a su vez provoca la proliferación celular de las células malignas de colon, es decir, la propagación del cáncer de colon. Por otro lado, la actividad anormalmente baja de la ruta también podría ser de interés, por ejemplo en el caso de la osteoporosis.

20 Las tecnologías para la adquisición de datos genómicos y proteómicos se han hecho fácilmente disponibles en entornos clínicos. Por ejemplo, las mediciones por microarreglos se emplean rutinariamente para evaluar los niveles de expresión génica, niveles de proteína, metilación, y así sucesivamente. La secuenciación automatizada de genes permite una identificación efectiva en costes de las variaciones genéticas en ADN y ARNm. La evaluación cuantitativa de los niveles de ARNm durante la secuenciación de genes promete ser otra herramienta clínica para evaluar los niveles de expresión génica.

30 A pesar de (o, tal vez, debido a) estos avances, la aplicación clínica de los análisis genómicos y proteómicos se enfrenta a un obstáculo sustancial: la sobrecarga de datos. Por ejemplo, el número de mutaciones identificables en una sola muestra clínica puede ser de cientos o miles. Muchas de estas mutaciones son denominadas mutaciones circunstanciales sin una contribución específica al crecimiento del cáncer, y solo unas pocas contribuyen al crecimiento del cáncer y a la evolución funcional, y éstas presentan las dianas para un tratamiento eficaz. Un microarreglo único puede generar niveles de expresión génica para decenas de miles de genes. El procesamiento de estas grandes cantidades de datos para identificar información clínicamente útil, como por ejemplo en la aplicación de la elección de la terapia adecuada, es difícil.

40 Un enfoque consiste en limitar el análisis a unas pocas pruebas canónicas o estandarizadas, como las pruebas aprobadas por la U.S. Food and Drug Administration (FDA). En este enfoque, se detecta un indicador específico o una combinación de indicadores (por ejemplo, mutaciones y/o niveles de expresión génica altos o bajos especificados) para probar como "positivo" para la condición de enfermedad indicada (por ejemplo, un tipo particular de cáncer). La prueba canónica es apoyada por estudios clínicos que han demostrado una fuerte correlación con la condición de la enfermedad o con la eficacia del tratamiento. Este enfoque es útil solo para aquellas condiciones clínicas para las que se ha desarrollado una prueba canónica, por ejemplo el diagnóstico específico de una enfermedad o la predicción de la respuesta a un fármaco en un tipo específico de cáncer en una etapa específica, y también es rígida, ya que solo es aplicable para las condiciones canónicas.

50 Otro enfoque se basa en la identificación de grupos funcionalmente relacionados de indicadores genómicos o proteómicos. Por ejemplo, la ruta Wnt comprende una cascada de reacciones proteómicas. Los componentes principales de esta cadena incluyen (pero no se limitan a) la unión de la proteína de señalización Wnt a un receptor de superficie rizado de la célula que provoca la activación de proteínas de la familia de proteínas Dsh que a su vez afectan el nivel de agentes de transcripción tales como  $\beta$ -catenina/TCF4 en el núcleo celular. Estos agentes de transcripción, a su vez, controlan la transcripción de moléculas de ARNm objetivo que a su vez se traducen en proteínas objetivo de la ruta Wnt. Los estudios clínicos han mostrado algunas correlaciones entre las proteínas reguladoras de la ruta Wnt y la actividad de la ruta Wnt.

55 Sin embargo, la aplicación de tales resultados del estudio clínico para el diagnóstico y la evaluación clínica de un paciente específico es difícil debido a la complejidad de las rutas de señalización, por ejemplo, el camino Wnt. Como un ejemplo simple, la medición del nivel de expresión de una proteína que está "corriente arriba" en la ruta Wnt puede no detectar el comportamiento anormal de una proteína que está "corriente abajo" en la ruta Wnt. Se cree que la ruta Wnt incluye numerosos mecanismos de retroalimentación y el concepto simplificado de "corriente arriba" y "corriente abajo" puede ser inaplicable para una porción sustancial de la ruta Wnt; más generalmente, el comportamiento anormal en una porción de la cascada de proteínas que comprende la ruta Wnt puede tener más o menos efecto sobre otras porciones de la cascada de proteínas y sobre la actividad de la ruta Wnt en su conjunto. Además, en algunos estudios clínicos, los niveles de expresión de proteína para las proteínas reguladoras de la cascada de señalización se evalúan midiendo los niveles de expresión de ARNm de los genes que codifican para las proteínas reguladoras. Esta es una

medida indirecta que puede no evaluar con precisión el nivel de expresión de la proteína reguladora, y casi nunca refleja la cantidad de proteínas activas (después de una modificación postraducción específica como la fosforilación).

5 Charles J. Vaske et al., " Interference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM", Bioinformatics, Vol. 26, No. 12, Junio de 2010, páginas i237 a i245 divulga un método para inferir actividades genéticas específicas del paciente que incorporan interacciones de rutas curadas entre genes. Un gen es modelado por un gráfico de factores como un conjunto de variables interconectadas que codifican la expresión y la actividad conocida de un gen y sus productos, permitiendo la incorporación de muchos tipos de datos ómicos como evidencia. El método predice el grado en que las actividades de una ruta (por ejemplo, estados génicos internos, interacciones o "salidas" de alto nivel) se alteran en el paciente utilizando la inferencia probabilística.

15 El principal problema subyacente a la presente invención era por lo tanto proporcionar métodos y medios adecuados para realizar análisis genómicos y, respectivamente, proteómicos. Los aspectos específicos del problema subyacente, así como otras objeciones relacionadas con la presente invención, se hacen evidentes cuando se estudia la descripción, los ejemplos proporcionados aquí y, en particular, cuando se estudian las reivindicaciones adjuntas.

La presente invención proporciona métodos y aparatos nuevos y mejorados como se describe en el presente documento.

20 De acuerdo con un aspecto principal de la presente invención, el problema anterior se resuelve mediante un método específico para evaluar la actividad de la ruta de señalización celular utilizando un modelo probabilístico de expresión génica objetivo, a saber, un procedimiento que comprende:

25 actividad de inferencia de una o más rutas de señalización celular en el tejido de un sujeto médico con base al menos en el nivel o niveles de expresión (en particular en el nivel de ARNm y/o proteína) de uno o más genes objetivo de las rutas de señalización celular medida en una muestra extraída del tejido del sujeto médico, en la que la inferencia comprende:

30 inferir la actividad de la(s) ruta(s) de señalización celular en el tejido del sujeto médico mediante la evaluación de al menos una porción de un modelo probabilístico, preferiblemente una red bayesiana, que representa las rutas de señalización celular para un conjunto de entradas que incluyen al menos el nivel o niveles de expresión de uno o más genes objetivo de las rutas de señalización celular medida en la muestra extraída del tejido del sujeto médico;

35 estimar un nivel en el tejido del sujeto médico de al menos un elemento de factor de transcripción (TF), controlando el al menos un elemento TF la transcripción del uno o más genes objetivo de las rutas de señalización celular, con base al menos en parte en probabilidades condicionales que relacionan al menos un elemento TF y el nivel o niveles de expresión del uno o más genes objetivo de las rutas de señalización celular medido en la muestra extraída del tejido del sujeto médico; y

40 inferir la actividad de la ruta de señalización celular con base en el nivel estimado en la muestra de tejido del elemento de factor de transcripción;

45 en el que la inferencia se realiza mediante un dispositivo de procesamiento digital que utiliza el modelo probabilístico de la(s) ruta(s) de señalización celular,

El "gen o genes objetivo" pueden ser "genes objetivo directos" y/o "genes objetivo indirectos" (como se describe en el presente documento).

50 Preferiblemente, la inferencia comprende estimar un nivel en el tejido del sujeto médico de al menos un elemento de factor de transcripción (TF) representado por un nodo TF del modelo probabilístico, controlando el elemento TF la transcripción del uno o más genes objetivo de la ruta de señalización celular, estando basada la estimación al menos en parte en probabilidades condicionales del modelo probabilístico que relaciona el nodo TF y los nodos en el modelo probabilístico que representa el uno o más genes objetivo de las rutas de señalización celular medida en la muestra extraída del tejido del sujeto médico.

55 El modelo probabilístico puede ser un modelo de red Bayesiana. De este modo, de acuerdo con una realización preferida, la inferencia se lleva a cabo utilizando una red Bayesiana que comprende nodos que representan información sobre las rutas de señalización y relaciones de probabilidad condicional entre nodos conectados de la red Bayesiana.

60 Las rutas de señalización celular pueden ser una ruta Wnt, una ruta ER (receptor de estrógeno), una ruta AR (receptor de andrógeno) y/o una ruta Hedgehog. Por lo tanto, de acuerdo con una realización preferida, las rutas de señalización celular comprenden una ruta Wnt, una ruta ER, una ruta AR y/o una ruta Hedgehog.

65

Los genes objetivo particularmente adecuados se describen en los siguientes pasajes de texto, así como en los ejemplos a continuación (véase, por ejemplo, las Tablas 1- 9).

5 De este modo, de acuerdo con una realización preferida, el gen o genes objetivo se seleccionan del grupo que comprende o consiste en los genes objetivo enumerados en la Tabla 1 o Tabla 6 (para la ruta Wnt), los genes objetivo enumerados en la Tabla 2, Tabla 5 o la Tabla 7 (para la ruta ER), los genes objetivo enumerados en la Tabla 3 o la Tabla 8 (para la ruta Hedgehog) y los genes objetivo enumerados en la Tabla 4 o la Tabla 9 (para la ruta AR).

10 Se prefiere particularmente un método en el que la inferencia comprende:

inferir la actividad de una ruta Wnt en el tejido del sujeto médico con base en al menos en los niveles de expresión de uno o más, preferiblemente al menos tres, genes de la ruta Wnt medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: KIAA1199, AXIN2, RNF43, TBX3, TDGF1, SOX9, ASCL2, IL8, SP5, ZNRF3, KLF6, CCND1, DEFA6 y FZD7.

15 Se prefiere además un método en el que la inferencia se basa adicionalmente en los niveles de expresión de al menos un gen objetivo de la ruta Wnt medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: NKD1, OAT, FAT1, LEF1, GLUL, REG1B, TCF7L2, COL18A1, BMP7, SLC1A2, ADRA2C, PPARG, DKK1, HNF1A y LECT2.

20 Particularmente preferido es un método en el que la inferencia (también) comprende:

inferir la actividad de una ruta ER en el tejido del sujeto médico con base en al menos en los niveles de expresión de uno o más, preferiblemente al menos tres genes objetivo de la ruta ER medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: CDH26, SGK3, PGR, GREB1, CA12, XBP1, CELSR2, WISP2, DSCAM, ERBB2, CTSD, TFF1 y NRIP1.

25 Se prefiere además un método, en el que la inferencia se basa adicionalmente en los niveles de expresión de al menos un gen objetivo de la ruta ER medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: AP1B1, ATP5J, COL18A1, COX7A2L, EBAG9, ESR1, HSPB1, IGFBP4, KRT19, MYC, NDUFV3, PISD, PRDM15, PTMA, RARA, SOD1 y TRIM25.

30 Un método en el que la inferencia (también) comprende

inferir la actividad de una ruta Hedgehog en el tejido del sujeto médico con base en al menos en los niveles de expresión de uno o más, preferiblemente al menos tres, genes objetivo de la ruta Hedgehog medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: GLI1, PTCH1, PTCH2, IGFBP6, SPP1, CCND2, FST, FOXL1, CFLAR, TSC22D1, RAB34, S100A9, S100A7, MYCN, FOXM1, GLI3, TCEA2, FYN y CTSL1, también se prefiere.

35 Se prefiere además un método, en el que la inferencia se basa adicionalmente en los niveles de expresión de al menos un gen objetivo de la ruta Hedgehog medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: BCL2, FOXA2, FOXF1, H19, HHIP, IL1R2, JAG2, JUP, MIF, MYLK, NKX2-2, NKX2-8, PITRM1 y TOM1.

40 Un método en el que la inferencia (también) comprende

inferir la actividad de una ruta AR en el tejido del sujeto médico con base en al menos en los niveles de expresión de uno o más, preferiblemente al menos tres, genes de la ruta AR medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: KLK2, PMEPA1, TMPRSS2, NKX3\_1, ABCC4, KLK3, FKBP5, ELL2, UGT2B15, DHCR24, PPAP2A, NDRG1, LRIG1, CREB3L4, LCP1, GUCY1A3, AR y EAF2, también se prefiere.

45 Se prefiere además un método en el que la inferencia se basa adicionalmente en los niveles de expresión de al menos un gen objetivo de la ruta AR medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende o consiste en: APP, NTS, PLAU, CDKN1A, DRG1, FGF8, IGF1, PRKACB, PTPN1, SGK1 y TACC2.

50 Se prefiere además un método, en el que el método comprende además:

determinar si las rutas de señalización celular está operando anormalmente en el tejido del sujeto médico con base en la actividad inferida de las rutas de señalización celular en el tejido del sujeto médico.

Otro aspecto de la presente invención se refiere a un método (como se describe en el presente documento), que comprende además:

60 Otro aspecto de la presente invención se refiere a un método (como se describe en el presente documento), que comprende además:

recomendar la prescripción de un fármaco para el sujeto médico que corrige el funcionamiento anormal de las rutas de señalización celular;

5 en el que la recomendación se realiza solo si se determina que las rutas de señalización celular está funcionando anormalmente en el tejido del sujeto médico con base en la actividad inferida de las rutas de señalización celular.

La presente invención se refiere también a un método (como se describe en el presente documento) que comprende:

10 inferir la actividad de una ruta Wnt en tejido de un sujeto médico con base al menos en los niveles de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta Wnt medida en una muestra extraída del tejido del sujeto médico

y/o

15 inferir la actividad de una ruta ER en tejido de un sujeto médico con base al menos en los niveles de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta ER medida en una muestra extraída del tejido del sujeto médico

y/o

20 inferir la actividad de una ruta Hedgehog en el tejido de un sujeto médico con base al menos en los niveles de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta Hedgehog medida en una muestra extraída del tejido del sujeto médico

25 y/o

inferir la actividad de una ruta AR en el tejido de un sujeto médico con base al menos en los niveles de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta AR medida en una muestra extraída del tejido del sujeto médico.

30

Preferiblemente,

35 el conjunto de genes objetivo de la ruta Wnt incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende o consiste en: KIAA1199, AXIN2, RNF43, TBX3, TDGF1, SOX9, ASCL2, IL8, SP5, ZNRF3, KLF6, CCND1, DEFA6 y FZD7,

y/o

40 el conjunto de genes objetivo de la ruta ER incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende o consiste en: CDH26, SGK3, PGR, GREB1, CA12, XBP1, CELSR2, WISP2, DSCAM, ERBB2, CTSD, TFF1 y NRIP1,

y/o

45 el conjunto de genes objetivo de la ruta Hedgehog incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende o que consiste en: GLI1, PTCH1, PTCH2, IGFBP6, SPP1, CCND2, FST, FOXL1, CFLAR, TSC22D1, RAB34, S100A9, S100A7, MYCN, FOXM1, GLI3, TCEA2, FYN y CTSL1,

y/o

50 el conjunto de genes objetivo de la ruta AR incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende o consiste en: KLK2, PMEPA1, TMPRSS2, NKX3-1, ABCC4, KLK3, FKBP5, ELL2, UGT2B15, DHCR24, PPAP2A, NDRG1, LRIG1, CREB3L4, LCP1, GUCY1A3, AR y EAF2.

55 Un método, en el que

el conjunto de genes objetivo de la ruta Wnt incluye además al menos un gen objetivo seleccionado del grupo que comprende o consiste en: NKD1, OAT, FAT1, LEF1, GLUL, REG1B, TCF7L2, COL18A1, BMP7, SLC1A2, ADRA2C, PPARG, DKK1, HNF1A y LECT2,

60

y/o

65 el conjunto de genes objetivo de la ruta ER incluye además al menos un gen objetivo seleccionado del grupo que comprende o consiste en: AP1B1, ATP5J, COL18A1, COX7A2L, EBAG9, ESR1, HSPB1, IGFBP4, KRT19, MYC, NDUFV3, PISD, PRDM15, PTMA, RARA, SOD1 y TRIM25,

y/o

el conjunto de genes objetivo de la ruta Hedgehog incluye además al menos un gen objetivo seleccionado del grupo que comprende o consiste en: BCL2, FOXA2, FOXF1, H19, HHIP, IL1R2, JAG2, JUP, MIF, MYLK, NKX2-2, NKX2-8, PITRM1 y TOM1,

y/o

se prefiere particularmente el conjunto de genes objetivo de la ruta AR que incluye además al menos un gen objetivo seleccionado del grupo que comprende o que consiste en: APP, NTS, PLAU, CDKN1A, DRG1, FGF8, IGF1, PRKACB, PTPN1, SGK1 y TACC2.

La muestra o muestras para usar de acuerdo con la presente invención pueden ser, por ejemplo, una muestra obtenida a partir de una lesión de mama, o de un colon de un sujeto médico conocido por o sospechoso de tener cáncer de colon, o de un hígado de un sujeto médico conocido por o sospechoso de tener cáncer de hígado, etc., preferiblemente mediante un procedimiento de biopsia u otro procedimiento de extracción de muestras. El tejido del que se extrae una muestra puede ser también tejido metastásico, por ejemplo tejido maligno (sospechoso) procedente de colon, mama, hígado u otro órgano que se ha diseminado fuera del colon, la mama, el hígado u otro órgano. En algunos casos, la muestra de tejido puede ser células tumorales circulantes, es decir, células tumorales que han entrado en el torrente sanguíneo y pueden extraerse como la muestra de tejido extraída usando técnicas de aislamiento adecuadas.

Otro aspecto divulgado de la presente invención se refiere al uso de un medio de almacenamiento no transitorio tal como se describe en el presente documento o a un programa informático como se describe aquí para el diagnóstico específico de una enfermedad o la predicción de la respuesta a un fármaco en un tipo específico de cáncer en una etapa específica.

De acuerdo con otro aspecto divulgado, un aparato comprende un procesador digital configurado para realizar un método según la invención tal como se describe en el presente documento.

De acuerdo con otro aspecto divulgado, un medio de almacenamiento no transitorio almacena instrucciones que son ejecutables mediante un dispositivo de procesamiento digital para llevar a cabo un método según la invención tal como se describe en el presente documento. El medio de almacenamiento no transitorio puede ser un medio de almacenamiento legible por ordenador, tal como un disco duro u otro medio de almacenamiento magnético, un disco óptico u otro medio de almacenamiento óptico, una memoria de acceso aleatorio (RAM), memoria de solo lectura (ROM), memoria instantánea u otro medio de almacenamiento electrónico, un servidor de red, etc. El dispositivo de procesamiento digital puede ser un dispositivo de mano (por ejemplo, un asistente de datos personales o un teléfono inteligente), un ordenador portátil, un ordenador de escritorio, un ordenador o dispositivo de tableta, un servidor de red remoto, etc.

De acuerdo con otro aspecto divulgado, un programa de ordenador comprende medios de código de programa para hacer que un dispositivo de procesamiento digital realice un método de acuerdo con la invención tal como se describe en el presente documento. El dispositivo de procesamiento digital puede ser un dispositivo de mano (por ejemplo, un asistente de datos personales o un teléfono inteligente), un ordenador portátil, un ordenador de escritorio, un ordenador o dispositivo de tableta, un servidor de red remoto, etc.

Una ventaja reside en un sistema de soporte de decisión clínica (CDS) que proporciona recomendaciones clínicas con base en el análisis probabilístico de una o más rutas de señalización celular, por ejemplo utilizando un modelo de red Bayesiana de una ruta Wnt, una ruta ER, una ruta AR y/o una ruta Hedgehog.

Otra ventaja reside en una evaluación mejorada de la actividad de la ruta de señalización celular que es menos susceptible al error.

Otra ventaja reside en proporcionar un sistema CDS que recomiende un tratamiento dirigido para la pérdida de la regulación de una ruta de señalización celular.

Otra ventaja reside en proporcionar un sistema CDS que está diseñado para detectar la pérdida de regulación para una ruta de señalización celular particular, tal como una ruta Wnt, una ruta ER, una ruta AR o una ruta Hedgehog y se adapta fácilmente para proporcionar recomendaciones para diferentes tipos de cáncer obtenidas por esa ruta de señalización celular particular.

La presente invención tal como se describe en el presente documento puede, también ventajosamente, por ejemplo, utilizarse en conexión con

- diagnóstico con base en la actividad predicha (inferida);

- pronóstico con base en la actividad predicha (inferida);

- prescripción de fármacos con base en la actividad predicha (inferida);
- predicción de la eficacia del fármaco con base en la actividad predicha (inferida);
- 5 - predicción de efectos adversos con base en la actividad predicha (inferida);
- control de la eficacia de los fármacos;
- desarrollo de fármacos;
- 10 - desarrollo del ensayo;
- investigación de rutas;
- 15 - estadificación del cáncer;
- inscripción del sujeto en un ensayo clínico con base en la actividad predicha (inferida);
- selección de la prueba posterior por realizar, y/o;
- 20 - selección de pruebas de diagnóstico complementarias.

25 Otras ventajas resultarán evidentes para los expertos en la materia tras la lectura y comprensión de las figuras adjuntas, de la siguiente descripción y, en particular, de la lectura de los ejemplos detallados proporcionados más adelante.

30 La Figura 1 muestra una red Bayesiana simple que representa parte de una ruta de señalización celular. La ruta de señalización celular está simbolizada por un complejo de factor de transcripción (TF) y los genes objetivo producidos como resultado de la presencia del complejo de factor de transcripción. La relación probabilística entre el elemento TF y un gen objetivo en caso de discretización binaria puede representarse mediante una tabla de probabilidad condicional como se representa en el diagrama.

35 La Figura 2 muestra una red Bayesiana ilustrativa que describe una ruta de señalización celular hipotética. En el diagrama se representan tanto las proteínas corriente arriba como los nodos de ARNm objetivo corriente abajo. Las proteínas corriente arriba sirven como entrada en el complejo de factor de transcripción, mientras que los ARNm objetivo son los nodos de salida del complejo del factor de transcripción.

40 La Figura 3 muestra un ejemplo ilustrativo de una representación de red Bayesiana de una única ruta de señalización celular con complejo de múltiples factores de transcripción o múltiples rutas de señalización celular con su propio complejo de factor de transcripción combinado en una red Bayesiana o una combinación de las mismas.

45 La Figura 4 muestra un ejemplo de una red Bayesiana que ilustra una representación simple de una ruta de señalización celular similar a la Figura 1. Ahora se han unido otros nodos para representar la traducción del ARNm objetivo en proteínas objetivo.

La Figura 5 muestra un ejemplo de una red Bayesiana que ilustra otra representación simple de una ruta de señalización celular. La ruta se representa utilizando el factor de transcripción complejo y sus niveles de proteína objetivo.

50 La Figura 6 muestra la red Bayesiana ilustrativa de la Figura 1 con una capa adicional de nodos que representan los conjuntos de sondas en un chip de microarreglos que conecta las intensidades de las sondas con los correspondientes niveles de ARNm objetivo.

55 La Figura 7 muestra un ejemplo ilustrativo de una variante de realización de la red Bayesiana de la Figura 1 que incluye nodos que representan variaciones en metilación y número de copias como ejemplos para nodos de información adicionales para, en este ejemplo particular, cualquiera de los niveles de ARNm objetivo incluidos.

60 La Figura 8 muestra una actividad de la ruta de Wnt predicha de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de colon (GSE20916).

La Figura 9 muestra una actividad de la ruta de Wnt predicha de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de colon (GSE4183).

65 La Figura 10 muestra una actividad de la ruta de Wnt predicha de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de colon (GSE15960).

- La Figura 11 muestra una actividad de la ruta de Wnt predicha de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de cáncer de mama (GSE12777).
- 5 La Figura 12 muestra una actividad de la ruta de Wnt predicha de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de cáncer de mama (GSE21653).
- La Figura 13 muestra una actividad de la ruta de Wnt pronosticada de la red Bayesiana y el método de centroide más cercano como se describe aquí en un conjunto de datos de muestras de cáncer de hígado (GSE9843).
- 10 La Figura 14 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana usando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe en el presente documento en un conjunto de datos de muestras de colon (GSE20916).
- 15 La Figura 15 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana usando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe aquí en un conjunto de datos de muestras de colon (GSE4183).
- La Figura 16 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana usando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe aquí en un conjunto de datos de muestras de colon (GSE15960).
- 20 La Figura 17 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana usando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe en el presente documento en un conjunto de datos de muestras de cáncer de mama (GSE12777).
- 25 La Figura 18 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana utilizando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe en el presente documento en un conjunto de datos de muestras de cáncer de hígado (GSE9843).
- 30 La Figura 19 muestra una actividad de la ruta de Wnt predicha utilizando una red Bayesiana utilizando los genes objetivo de la lista de curado de evidencia en comparación con los genes objetivo de la lista de literatura amplia como se describe en el presente documento en un conjunto de datos de muestras de meduloblastoma (GSE10327).
- 35 La Figura 20 muestra esquemáticamente un sistema de soporte de decisión clínica (CDS) configurado para evaluar una o más rutas de señalización celular tal como se describe aquí (ejemplar mostrado para la ruta Wnt).
- La Figura 21 muestra una actividad prevista de la ruta de Wnt en muestras de colon de GSE4183.
- 40 La Figura 22 muestra una actividad prevista de la ruta Wnt en muestras de meduloblastoma de GSE10327.
- La Figura 23 muestra una actividad prevista de la ruta Wnt en muestras de cáncer de hígado de GSE9843.
- La Figura 24 muestra una actividad prevista de la ruta Wnt en líneas celulares de cáncer de mama de GSE12777.
- 45 La Figura 25 muestra una actividad prevista de la ruta ER en las líneas celulares de cáncer de mama de GSE12777.
- La Figura 26 muestra una actividad prevista de la ruta ER en muestras de cáncer de mama de GSE12276.
- 50 La Figura 27 muestra una actividad prevista de la ruta ER en las líneas celulares de cáncer de GSE36133.
- La Figura 28 muestra una actividad de la ruta prevista de Hedgehog en líneas celulares de cáncer de GSE34211.
- La Figura 29 muestra una actividad prevista de la ruta Hedgehog en muestras de meduloblastoma de GSE10327.
- 55 La Figura 30 muestra una actividad prevista de la ruta Hedgehog en muestras de cáncer de mama de GSE12276.
- La Figura 31 muestra una actividad de ruta ER esperada en MCF7 y líneas de células resistentes a tamoxifeno de GSE21618.
- 60 La Figura 32 muestra una actividad predicha de la ruta ER en una serie temporal de muestras de células MCF7 estimuladas con estrógenos de GSE11324.
- La Figura 33 muestra la actividad de la ruta Wnt, ER y Hedgehog en muestras de luminal A de GSE12276.
- 65 La Figura 34 muestra la actividad de las rutas Wnt, ER y Hedgehog en muestras basales de GSE12276.

La Figura 35 muestra una actividad prevista de la ruta Wnt en muestras de colon de GSE20916.

La Figura 36 muestra una actividad de la ruta ER esperada en líneas celulares MCF7 estimuladas con estrógeno (E2) o un control negativo (EtOH) (GSE9253).

La Figura 37 muestra las curvas de supervivencia de Kaplan-Meier de los pacientes del conjunto de datos GSE12276 agrupados de acuerdo con la actividad de la ruta.

La Figura 38 muestra una actividad de ruta AR predicha en líneas celulares LNCaP tratadas con diferentes regímenes de tratamiento de GSE7708.

La Figura 39 muestra una actividad predicha de ruta AR en muestras de cáncer de próstata de GSE17951.

La Figura 40 muestra una actividad de ruta AR predicha en muestras de cáncer de mama de GSE12276.

La Figura 41 muestra una actividad de ruta AR predicha en el conjunto de datos GSE36133 que contiene muestras de líneas celulares que representan diversos tipos de cáncer.

La Figura 42 muestra una actividad de ruta AR predicha en el conjunto de datos GSE34211 que contiene muestras de líneas celulares que representan diversos tipos de cáncer.

Los siguientes ejemplos solamente ilustran métodos y aspectos particularmente preferidos seleccionados en conexión con los mismos. La enseñanza proporcionada en los mismos puede ser usada para construir varios ensayos y/o kits, por ejemplo para detectar, predecir y/o diagnosticar la actividad anormal de una o más rutas de señalización celular. Además, al utilizar métodos como se describen aquí, se puede guiar ventajosamente la prescripción de fármacos, se puede predecir el fármaco y monitorizar la eficacia (y/o efectos adversos) del fármaco, se puede predecir y controlar la resistencia a fármacos, por ejemplo para seleccionar las pruebas posteriores por realizar (como una prueba de diagnóstico complementaria). Los siguientes ejemplos no deben interpretarse como limitativos del alcance de la presente invención.

#### Ejemplo 1: Construcción de la red Bayesiana

Tal como se describe en el presente documento, construyendo un modelo probabilístico (por ejemplo, el modelo Bayesiano ilustrativo mostrado en la Figura 6) e incorporando relaciones probabilísticas condicionales entre los niveles de expresión de un número de diferentes genes objetivo y la actividad de la ruta de señalización celular, tal modelo puede usarse para determinar la actividad de la ruta de señalización celular con un alto grado de precisión. Además, el modelo probabilístico puede actualizarse fácilmente para incorporar conocimientos adicionales obtenidos en estudios clínicos posteriores, ajustando las probabilidades condicionales y/o añadiendo nuevos nodos al modelo para representar fuentes de información adicionales. De esta manera, el modelo probabilístico puede ser actualizado como apropiado para incorporar el conocimiento médico más reciente.

Uno de los modelos de red Bayesiana más simples para representar una ruta de señalización celular sería un modelo de dos niveles que incluyera el elemento de factor de transcripción y los genes objetivo asociados (véase la Figura 1). El elemento complejo del factor de transcripción es una representación del nivel del complejo del factor de transcripción. El nivel de proteína del elemento de factor de transcripción está conectado a un número de niveles de ARNm de los genes objetivo del factor de transcripción (en esta red Bayesiana ejemplar solo se representan tres genes objetivo que se sabe que se expresan en el tejido en caso de que el factor de transcripción esté disponible). Debe entenderse que muchos, la mayoría o todos los genes objetivo de la ruta (en el caso de las rutas Wnt, ER, Hedgehog y AR particularmente los genes objetivo mencionados en Tabla 1, Tabla 2, Tabla 3 y Tabla 4 respectivamente) son regulados de manera análoga por el elemento TF. Las relaciones entre el nivel del elemento TF y los niveles de ARNm de los genes objetivo se modelan en la red Bayesiana por los bordes. Para cada uno de los genes objetivo, una distribución probabilística condicional específica cómo el nivel de ARNm del gen depende del nivel del elemento TF.

Los niveles del elemento TF y los genes objetivo pueden estar representados de diversas maneras. Una opción es usar una discretización binaria, en estados "ausentes" y "presentes" para el elemento TF, y "abajo" y "arriba" para el nivel de ARNm de un gen objetivo (ver Figura 1). La relación probabilística entre el elemento TF y un gen objetivo puede entonces representarse mediante una tabla de probabilidad condicional (como se indica en la misma Figura).

En lugar de una discretización binaria, los niveles también pueden representarse como valores de nivel continuo, o como valores cuantificados que tienen tres o más niveles de cuantificación (por ejemplo, "abajo", "normal" y "arriba" para genes objetivo).

La ilustración anterior de una red Bayesiana simple es solo una realización ilustrativa del modelo de red Bayesiana (Figura 1). En general, un modelo de red Bayesiana comprende un grafo acíclico dirigido que comprende nodos conectados por bordes. Cada nodo representa un elemento de información perteneciente a la ruta a mano (o, más

generalmente, a la ruta de señalización celular). Cada uno de los nodos de los elementos de la ruta representa un elemento genómico o proteómico de la ruta de señalización celular. A modo de ejemplo ilustrativo, un nodo de elemento de ruta puede representar uno de entre, pero no restringiéndose a: una proteína, un complejo de proteína, una molécula de ARNm transcrita a partir de un gen objetivo de la ruta de señalización celular, un gen metilado, una proteína fosforilada, un complejo de proteína fosforilada, o así sucesivamente. Como se describe más adelante en este documento, pueden incluirse en la red Bayesiana otros tipos de nodos, pero no se limitan a los ejemplos dados, para representar otros tipos de información tales como elementos de referencia de medición específicos, ocurrencias de variación génica, etc.

Típicamente, se añaden niveles adicionales "corriente arriba" que representan las proteínas reguladoras (en estado activo o inactivo) si el conocimiento del nivel de dicha proteína pudiera ser probatorio para determinar la recomendación de soporte de decisión clínica. Por ejemplo, podría ser útil la inclusión de las proteínas elementales en el factor de transcripción o en las proteínas esenciales antes del factor de transcripción en la red Bayesiana (véase la Figura 2) si se dispone de un fármaco que específicamente apunte a dichas proteínas y no a la ruta como un todo. Se cree que el factor de transcripción (TF) es un complejo proteico (es decir, una combinación de proteínas unidas entre sí en una estructura específica que realiza la función de regular la transcripción a partir de los genes objetivo) en la mayoría de las rutas de señalización. Para otras rutas, el elemento TF puede ser una sola proteína. Además, las rutas de señalización pueden ejercer su actividad a través de más de un factor de transcripción, resultando en una más compleja red Bayesiana con múltiples factores de transcripción de alimentación en los genes objetivo (véase la Figura 3 para una ilustración hipotética de múltiples elementos de factor de transcripción que influyen en la transcripción del gen objetivo). Esta red Bayesiana de múltiples factores de transcripción puede ser también el resultado de una combinación de rutas combinadas en una red Bayesiana.

También se pueden incluir nodos de información adicionales más abajo de los genes objetivo en la red Bayesiana. Un ejemplo ilustrativo de esto es la traducción del ARNm del gen objetivo en proteínas (Figura 4) o nodos de nivel de proteína del gen objetivo como nodo sustituto del nivel del ARNm del gen objetivo (Figura 5). Las moléculas de ARNm del gen objetivo se traducen por interacción con moléculas de ribosomas para formar proteínas que corresponden a las moléculas de ARNm y que corresponden a los genes objetivo. Esta es la expresión de los genes objetivo en el nivel de proteína. La medición del nivel de proteína mediante, pero sin limitarse a, por ejemplo espectrometría de masas, inmunohistoquímica, técnicas de electroforesis en gel, puede actuar como evidencia para estos niveles de proteína objetivo.

El nivel de expresión de un gen objetivo se puede calcular con base en la intensidad medida de conjuntos de sondas correspondientes de un microarreglo, por ejemplo, promediando o por otros medios de otras técnicas (por ejemplo, secuenciación de ARN). En algunas realizaciones este cómputo se integra en la red Bayesiana, extendiendo la red Bayesiana con un nodo para cada conjunto de sondas que se utiliza e incluyendo un recorrido de borde a cada uno de estos nodos de "medición" desde el nodo de gen objetivo correspondiente, como se describe en el presente documento con referencia a la Figura 6.

Opcionalmente, el modelo probabilístico también puede incorporar información genómica adicional, tal como información sobre mutaciones, variaciones de número de copias, expresión génica, metilación, información de translocación, etc., que cambian secuencias genómicas que están relacionadas con la cascada de señalización de la ruta para inferir la actividad de la ruta y para localizar el defecto en la ruta Wnt que provoca el funcionamiento aberrante (ya sea activación o inactividad), como se describe por referencia ilustrativa a la Figura 7 para el caso ilustrativo de la metilación y el número de copias de datos. Sin embargo, debe entenderse que otros tipos de información con respecto al gen objetivo se traducen análogamente en nodos de información. Tal información genómica puede estar disponible a través, pero no limitándose a, secuenciación de ARN y análisis de SNP.

Además, debe entenderse que mientras que los ejemplos tal como se describen más adelante en este documento pertenecen a la ruta Wnt, ER, AR y Hedgehog se proporcionan como ejemplos ilustrativos, los enfoques para el análisis de la ruta de señalización celular descritos aquí se aplican fácilmente a otras rutas de señalización celular además de estas (por ejemplo, las rutas de señalización celular Notch, HER2/PI3K, TGFbeta, EGF, VEGF y TNF-NFkappaB), y rutas de señalización intracelular con receptores dentro de la célula (por ejemplo, rutas de señalización celular de progesterona, ácido retinoico y vitamina D).

Ejemplo 2: Comparación de métodos de aprendizaje de máquina

Aquí se compara el rendimiento de dos tipos de técnicas de aprendizaje de máquina entre sí tomando la ruta Wnt como ejemplo: la predicción de la actividad Wnt mediante el método centroide más próximo se compara con el método de elección según la presente invención, que por ejemplo utiliza una red Bayesiana.

Como se discutió anteriormente, el enfoque de la red Bayesiana se seleccionó con base en sus ventajas que residen en el enfoque probabilístico, pudiendo incorporar la información disponible en forma "blanda", por ejemplo porcentajes de sujetos de estudio que muestran características probatorias, y en forma "dura", usando relaciones probabilísticas condicionales. Además, el modelo probabilístico también permite incorporar información a partir del conocimiento

parcial (más que integral) de la ruta de señalización celular subyacente, también mediante el uso de tablas de probabilidad condicional.

Aquí se demuestra que los inventores añadieron valor en la forma en que incluían propiedades biológicas conocidas y la disponibilidad de pruebas blandas utilizando una red Bayesiana en comparación con otros métodos de aprendizaje de máquina, por ejemplo clasificación centroide más cercana, un método bien conocido. La clasificación centroide más cercana es un método de aprendizaje de máquina donde para cada clase de muestras de entrenamiento se calcula un perfil medio (= centroide), y luego, para que una muestra sea clasificada, la etiqueta se predice en base al centroide más cercano (la etiqueta centroide más cercana es por tanto el resultado de la predicción). Los dos centroides se calculan en la misma lista de conjuntos de sonda utilizados en la red Bayesiana, y para el "Wnt en" y "Wnt fuera" del centroide se basan en las muestras de adenoma y las muestras normales de colon, respectivamente, de los mismos datos procesados fRMA de GSE8671. La relación  $\log_2$  de las dos distancias euclidianas entre una muestra y los dos centroides se utilizó posteriormente para clasificar las muestras de diversos conjuntos de datos para inferir la clasificación de las muestras. Esto significa que una relación  $\log_2$  de 0 corresponde a una distancia igual de la muestra a los dos centroides, un valor  $> 0$  corresponde a una muestra clasificada como señalización Wnt activa mientras que un valor  $< 0$  corresponde a una muestra identificada por tener una ruta de señalización Wnt inactiva.

La red Bayesiana se construyó similar a la Figura 6 y al procedimiento descrito en el presente documento. De forma similar a esta descripción de la red Bayesiana Wnt, las tablas de probabilidad condicional de los bordes entre conjuntos de sonda y sus respectivos genes fueron entrenados utilizando datos procesados fRMA de 32 muestras de colon normal y 32 muestras de adenoma del conjunto de datos GSE8671 de Gene Expression Omnibus (accesible en <http://www.ncbi.nlm.nih.gov/geo/>, visitado por última vez el 13 de julio de 2011). La red Bayesiana entrenada fue entonces probada en diversos conjuntos de datos para inferir la probabilidad P (Wnt habilitada) de que la ruta Wnt está "habilitada", es decir, activa, lo que se toma igual a la probabilidad deducida de que el complejo de transcripción de la ruta Wnt está "presente".

La red Bayesiana entrenada y el modelo de centroide más cercano se ensayaron entonces en varios conjuntos de datos de microarreglos procesados con fRMA para inferir la probabilidad de que la ruta de Wnt esté "habilitada", medida por P (Wnt Habilitada) y relación  $\log_2$  de las distancias. Los resúmenes de los resultados de la red Bayesiana y el modelo de centroide más cercano se muestran en las Figuras 8 a 13. El lector debe tener en cuenta que las métricas de salida de los dos métodos no son una relación uno a uno, sin embargo el signo y la magnitud relativa de las métricas de salida dentro de un método son comparables.

La gran mayoría de las muestras de colon (cáncer) (GSE20916, GSE4183) se clasifican por igual entre la ruta Wnt activa e inactiva, excepto GSE15960 que tenía una fracción elevada de muestras negativas clasificadas erróneamente en el método centroide más cercano (falsos negativos). Esta percepción de una mayor fracción de falsos negativos se mantiene en los otros tipos de cáncer también. Esto es especialmente cierto para muestras de cáncer de mama (GSE12777, GSE21653) y cáncer de hígado (GSE9843); salvo algunas excepciones, se prevé que todas las muestras tengan una ruta de Wnt inactiva que se sabe que es incorrecta en el caso de muestras de cáncer de mama de tipo basal y de cáncer de hígado CTNNB1. En algunos casos, evidente en, por ejemplo, GSE15960, la clasificación se podría corregir bajando y aumentando el umbral de la clasificación centroide más cercana. La idea detrás de esto sería que el umbral de la actividad Wnt podría ser alterado en diferentes tipos de tejido. Sin embargo, esto implicaría una formación adicional del método centroide más cercano para que sea aplicable a otros tipos de tejidos. Una de las fortalezas del modelo de la red Bayesiana es que este entrenamiento específico del tejido no es necesario ya que se establece que no es específico en cuanto al tipo de tejido.

### Ejemplo 3: Selección de genes objetivo

Un factor de transcripción (TF) es un complejo de proteínas (es decir, una combinación de proteínas unidas entre sí en una estructura específica) o una proteína que es capaz de regular la transcripción a partir de genes objetivo mediante la unión a secuencias de ADN específicas, controlando por lo tanto la transcripción de información genética de ADN a ARNm. El ARNm producido directamente debido a esta acción del complejo de transcripción se denomina en el presente documento como un "gen objetivo directo". La activación de la ruta también puede resultar en más transcripción de genes secundarios, denominados "genes objetivo indirectos". En lo que sigue, se prefieren los modelos de red Bayesiana (como modelos probabilísticos ejemplares) que comprenden o consisten en genes objetivo directos, como enlaces directos entre la actividad de la ruta y el nivel de ARNm, sin embargo la distinción entre genes objetivo directos e indirectos no es siempre evidente. Aquí se presenta un método para seleccionar genes objetivo directos usando una función de puntuación con base en datos de literatura disponible. Sin embargo, la selección accidental de genes objetivo indirectos no puede descartarse debido a la información limitada y a las variaciones e incertidumbres biológicas.

Se seleccionaron genes de ARNm de ruta específica a partir de la bibliografía científica, utilizando un sistema de clasificación en el que se daba una calificación a la evidencia científica de un gen objetivo específico, dependiendo del tipo de experimentos científicos en los que se acumulaba la evidencia. Aunque algunas pruebas experimentales son meramente indicativas de que un gen es un gen objetivo, como por ejemplo un ARNm que aumenta en un microarreglo de un embrión en el que se sabe que la ruta Hedgehog está activa, otras pruebas pueden ser muy fuertes,

como la combinación de un sitio de unión del factor de transcripción de la ruta identificada y la recuperación de este sitio en un ensayo de inmunoprecipitación de cromatina (ChIP) después de la estimulación de la ruta específica en la célula y el aumento de ARNm después de la estimulación específica de la ruta en una línea celular.

5 En la literatura científica pueden identificarse varios tipos de experimentos para encontrar genes objetivo específicos de la ruta de acceso:

10 1. Experimentos ChIP en los que se muestra la unión directa de un factor de transcripción de ruta a su sitio de unión en el genoma. Ejemplo: Mediante el uso de la tecnología de cromatinainmunoprecipitación (ChIP), se identificaron posteriores sitios de unión a factores de transcripción TCF4 funcionales en el ADN de las líneas celulares de colon con y sin ruta Wnt activa, como un subconjunto de los sitios de unión reconocidos puramente con base en la secuencia de nucleótidos. La funcionalidad putativa se identificó como la evidencia derivada de ChIP de que se encontró que el factor de transcripción se unía al sitio de unión al ADN.

15 2. Ensayos de Desplazamiento de Movilidad Electroforética (EMSA) que muestran la unión *in vitro* de un factor de transcripción a un fragmento de ADN que contiene la secuencia de unión. En comparación con la evidencia con base ChIP, la evidencia con base en EMSA es menos fuerte, ya que no puede trasladarse a la situación *in vivo*.

20 3. Estimulación de la ruta y medición de los perfiles de ARNm en un microarreglo o utilizando secuenciación de ARN, utilizando líneas celulares inducibles por las rutas y la medición de los perfiles ARNm medidos en varios puntos de tiempo después de la inducción en presencia de cicloheximida, lo que inhibe la traducción a la proteína, suponiéndose así que los ARNm inducidos son genes objetivo directos.

25 4. Similar a 3, pero utilizando la PCR cuantitativa para medir las cantidades de ARNm.

30 5. Identificación de los sitios de unión al factor de transcripción en el genoma utilizando un enfoque bioinformático. Ejemplo para la ruta Wnt: Usando la secuencia conocida de unión al ADN del factor de transcripción de la catenina TCF4-beta, se ejecutó un programa de software sobre la secuencia del genoma humano, y se identificaron sitios potenciales de unión, tanto en regiones promotoras de gen como en otras regiones genómicas.

6. Semejante a 3, solo en ausencia de cicloheximida.

7. Similar a 4, solo en ausencia de cicloheximida.

35 8. Perfil de expresión de ARNm de muestras de tejidos o células específicas de las que se sabe que la ruta es activa, sin embargo en ausencia de la condición de control negativo apropiada.

40 En la forma más sencilla se puede dar a cada ARNm objetivo potencial 1 punto para cada uno de estos enfoques experimentales en los que se identificó el ARNm objetivo.

Alternativamente, los puntos se pueden dar de forma incremental, es decir, una tecnología 1 punto, la segunda tecnología añade un segundo punto, y así sucesivamente. Usando esta estrategia relativamente clasificadora, se puede hacer una lista más confiable de los genes objetivo.

45 Alternativamente, la clasificación de otra manera puede usarse para identificar los genes objetivo que tienen más probabilidades de ser genes objetivo directos, proporcionando un mayor número de puntos a la tecnología que proporciona la mayor evidencia de un gen objetivo directo *in vivo*; en la lista anterior significaría 8 puntos para el enfoque experimental 1), 7 a 2), y bajando a un punto para la metodología experimental 8. Tal lista puede denominarse "lista de genes objetivo general".

50 A pesar de las variaciones e incertidumbres biológicas, los inventores asumieron que los genes objetivo directos son los que tienen más probabilidades de ser inducidos de una manera independiente de los tejidos. Una lista de estos genes objetivo puede ser llamada "lista de genes objetivo curada con evidencia". Estas listas de objetivos curados se han utilizado para construir modelos computacionales que se pueden aplicar a muestras procedentes de diferentes fuentes de tejido.

55 La "lista general de genes objetivo" probablemente contiene genes que son más específicos de tejido, y pueden usarse potencialmente para optimizar y aumentar la sensibilidad y especificidad del modelo para aplicación en muestras de un tejido específico, como muestras de cáncer de mama.

60 A continuación se ilustra un ejemplo de cómo se construyó específicamente la selección de una lista de genes objetivo curada con evidencia para la ruta ER.

Con el fin de seleccionar los genes objetivo ER usados como entrada para el "modelo", se usaron los tres criterios siguientes:

1. La región promotora/potenciadora de gen contiene un motivo de elemento de respuesta a estrógenos (ERE):

a. Se debe demostrar que el motivo ERE responde al estrógeno, por ejemplo, mediante un ensayo de transfección transitoria en el que el motivo ERE específico está unido a un gen informador y

b. La presencia del motivo ERE debe confirmarse, por ejemplo, mediante un análisis de motivo enriquecido de la región promotora/potenciadora del gen.

2. ER (diferencialmente) se une *in vivo* a la región promotora/potenciadora del gen en cuestión, demostrado, por ejemplo, mediante un experimento de CHIP/CHIP o un ensayo de inmunoprecipitación de cromatina:

a. Se ha demostrado que ER se une a la región promotora/potenciadora del gen cuando la ruta ER es activa, y

b. (preferiblemente) no se une (o se une débilmente) a la región promotora/potenciadora del gen si la ruta ER no está activa.

3. El gen es transcrito diferencialmente cuando la ruta ER es activa, demostrada por, por ejemplo,

a. enriquecimiento del ARNm del gen en cuestión mediante PCR en tiempo real, o experimento con microarreglos, o

b. la demostración de que el ARN Pol II se une a la región promotora del gen a través de un ensayo de inmunoprecipitación.

La selección se realizó mediante la definición como genes ER objetivo de los genes para los que se reunieron pruebas experimentales suficientes y bien documentadas que demostraban que los tres criterios mencionados anteriormente se cumplieron. Un experimento adecuado para recopilar evidencia de unión diferencial ER es comparar los resultados de, por ejemplo, un experimento CHIP/CHIP en una línea celular cancerosa que responde a estrógeno (por ejemplo, la línea celular MCF-7), cuando se expone o no se expone a estrógeno. Lo mismo ocurre con la recolección de pruebas de la transcripción de ARNm.

La metodología anterior discute el enfoque genérico y un ejemplo más específico del procedimiento de selección de genes objetivo que se ha empleado para seleccionar un número de genes objetivo con base en la evidencia encontrada usando la metodología mencionado anteriormente. En la Tabla 1, Tabla 2, Tabla 3 y Tabla 4 se muestran las listas de genes objetivo utilizados en los modelos de red Bayesiana para rutas ejemplares, a saber, las rutas Wnt, ER, Hedgehog y AR.

Los genes objetivo de la ruta ER usados para el modelo de red Bayesiana de la ruta ER descrita en este documento (mostrada en la Tabla 2) contienen una selección de genes objetivo con base en su puntuación de evidencia bibliográfica; Solo se añadieron a esta breve lista los genes objetivo con las puntuaciones de evidencia más altas (genes objetivo preferidos de acuerdo con la invención). En la Tabla 5 se muestra la lista completa de genes objetivo ER, incluyendo también aquellos genes con una puntuación de evidencia menor.

Se llevó a cabo otra subselección o clasificación de los genes objetivo de las rutas Wnt, ER, Hedgehog y AR mostrados en la Tabla 1, Tabla 2, Tabla 3 y Tabla 4 con base en una combinación del puntaje de evidencia de la literatura y las relaciones de probabilidades calculadas usando las tablas de probabilidad condicional entrenadas que enlazan los nodos de sondas con los nodos de genes objetivo correspondientes. La relación de probabilidades es una evaluación de la importancia del gen objetivo en la actividad de inferencia de las rutas. En general, se espera que sea probable que el nivel de expresión de un gen objetivo con una relación de probabilidades mayor que sea más informativo en cuanto a la actividad global de la ruta en comparación con genes objetivo con menores relaciones de probabilidades. Sin embargo, debido a la complejidad de las rutas de señalización celular debe entenderse que pueden existir interrelaciones más complejas entre los genes objetivo y la actividad de la ruta -por ejemplo, considerando los niveles de expresión de diversas combinaciones de genes objetivo con relaciones de probabilidad bajas quizás más probables que considerando los genes objetivo con mayores relaciones de probabilidades en aislamiento. En los modelos Wnt, ER, Hedgehog y AR, se ha encontrado que los genes objetivo mostrados en la Tabla 6, Tabla 7, Tabla 8 y Tabla 9 son de naturaleza probatoria superior para predecir la ruta Wnt, ER, Hedgehog y las actividades de ruta AR en comparación con los genes objetivo de menor rango (por lo tanto, los genes objetivo mostrados en las Tablas 6 a 9 son particularmente preferidos de acuerdo con la presente invención). No obstante, dada la relativa facilidad con la que la tecnología de adquisición tal como la de microarreglos puede adquirir niveles de expresión para grandes conjuntos de genes, se contempla utilizar algunos o todos los genes objetivo de la Tabla 6, Tabla 7, Tabla 8 y Tabla 9 y opcionalmente de forma adicional, usar uno, dos, algunos o todos los genes objetivo adicionales de los rangos mostrados en la Tabla 1, Tabla 2, Tabla 3 y Tabla 4, en el modelo Bayesiano como se representa en la Figura 6.

ES 2 700 617 T3

Tabla 1. Lista curada con evidencia de los genes objetivo de la ruta Wnt utilizados en la red bayesiana y conjuntos de sondas asociados usados para medir el nivel de expresión ARNm de los genes objetivo (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	Conjunto de sondas	#	Genes objetivo	Conjunto de sondas	#
ADRA2C	206128_at	4	HNF1A	210515_at	102
ASCL2	207607_at	10		216930_at	
	229215_at		IL8	202859_x_at	110
AXIN2	222695_s_at	13		211506_s_at	
	222696_at		KIAA1199	1554685_a_at	119
	224176_s_at			212942_s_at	
	224498_x_at		KLF6	1555832_s_at	121
BMP7	209590_at	17		208960_s_at	
	209591_s_at			208961_s_at	
	211259_s_at			211610_at	
	211260_at			224606_at	
CCND1	208711_s_at	27	LECT2	207409_at	129
	208712_at		LEF1	210948_s_at	130
	214019_at			221557_s_at	
CD44	1557905_s_at	30		221558_s_at	
	1565868_at		LGR5	210393_at	131
	204489_s_at			213880_at	
	204490_s_at		MYC	202431_s_at	142
	209835_x_at			244089_at	
	210916_s_at		NKD1	1553115_at	150
	212014_x_at			229481_at	
	212063_at			232203_at	
	216056_at		OAT	201599_at	157
	217523_at		PPARG	208510_s_at	173
	229221_at		REG1B	205886_at	184
	234411_x_at		RNF43	218704_at	189
	234418_x_at		SLC1A2	1558009_at	200
COL18A1	209081_s_at	40		1558010_s_at	
	209082_s_at			208389_s_at	
DEFA6	207814_at	52		225491_at	
DKK1	204602_at	54	SOX9	202935_s_at	209
EPHB2	209588_at	67		202936_s_at	
	209589_s_at		SP5	235845_at	210
	210651_s_at		TBX3	219682_s_at	215
	211165_x_at			222917_s_at	
EPHB3	1438_at	68		225544_at	
	204600_at			229576_s_at	
FAT1	201579_at	72	TCF7L2	212759_s_at	219
FZD7	203705_s_at	90		212761_at	
	203706_s_at			212762_s_at	
GLUL	200648_s_at	95		216035_x_at	
	215001_s_at			216037_x_at	
	217202_s_at			216511_s_at	
	217203_at			236094_at	
	242281_at		TDGF1	206286_s_at	220
			ZNRF3	226360_at	248

5

Tabla 2. Lista curada con evidencia de los genes objetivo de la ruta ER utilizados en la red bayesiana y los conjuntos de sonda asociados usados para medir el nivel de expresión ARNm de los genes objetivo (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	Conjunto de sondas	#	Genes objetivo	Conjunto de sondas	#
AP1B1	205423_at	5	RARA	1565358_at	183
ATP5J	202325_s_at	12		203749_s_at	
COL18A1	209081_s_at	40		203750_s_at	
	209082_s_at			211605_s_at	
COX7A2L	201256_at	41		216300_x_at	
CTSD	200766_at	46	SOD1	200642_at	205
DSCAM	211484_s_at	59	TFF1	205009_at	221

ES 2 700 617 T3

	237268 at		TRIM25	206911 at	230
	240218 at			224806 at	
EBAG9	204274 at	61	XBP1	200670 at	244
	204278 s at			242021 at	
ESR1	205225 at	70	GREB1	205862 at	97
	211233 x at			210562 at	
	211234 x at			210855 at	
	211235 s at		IGFBP4	201508 at	106
	211627 x at		MYC	202431 s at	142
	215551 at			244089 at	
	215552 s at		SGK3	227627 at	196
	217163 at			220038 at	
	217190 x at		WISP2	205792 at	241
	207672 at		ERBB2	210930 s at	69
HSPB1	201841 s at	103		216836 s at	
KRT19	201650 at	124		234354 x at	
	228491 at		CA12	203963 at	22
NDUFV3	226209 at	148		204508 s at	
	226616 s at			204509 at	
NRIP1	202599 s at	154		210735 s at	
	202600 s at			214164 x at	
PGR	208305 at	162		215867 x at	
	228554 at			241230 at	
PISD	202392 s at	164	CDH26	232306 at	32
PRDM15	230553 at	174		233391 at	
	230777 s at			233662 at	
	231931 at			233663 s at	
	234524 at		CELSR2	204029 at	36
	236061 at			36499 at	
PTMA	200772 x at	179			
	200773 x at				
	208549 x at				
	211921 x at				

Tabla 3. Lista curada con evidencia de los genes objetivo de la ruta Hedgehog utilizados en la red bayesiana y los conjuntos de sonda asociados usados para medir el nivel de expresión ARNm de los genes objetivo (# = número de secuencia en la lista de secuencias adjunta).

5

Genes objetivo	Conjunto de sondas	#	Genes objetivo	Conjunto de sondas	#
GLI1	206646 at	93	CTSL1	202087 s at	47
PTCH1	1555520 at	177	TCEA2	203919 at	216
	208522 s at			238173 at	
	209815 at			241428 x at	
	209816 at		MYLK	1563466 at	145
	238754 at			1568770 at	
PTCH2	221292 at	178		1569956 at	
HHIP	1556037 s at	101		202555 s at	
	223775 at			224823 at	
	230135 at		FYN	1559101 at	88
	237466 s at			210105 s at	
SPP1	1568574 x at	212		212486 s at	
	209875 s at			216033 s at	
TSC22D1	215111 s at	232	PITRM1	205273 s at	165
	235315 at			239378 at	
	243133 at		CFLAR	208485 x at	37
	239123 at			209508 x at	
CCND2	200951 s at	28		209939 x at	
	200952 s at			210563 x at	
	200953 s at			210564 x at	
	231259 s at			211316 x at	
H19	224646 x at	253		211317 s at	
	224997 x at			211862 x at	
IGFBP6	203851 at	107		214486 x at	

ES 2 700 617 T3

TOM1	202807 s at	229		214618 at	
JUP	201015 s at	117		217654 at	
FOXA2	210103 s at	82		235427 at	
	214312 at			237367 x at	
	40284 at			239629 at	
MYCN	209756 s at	144		224261 at	
	209757 s at		IL1R2	205403 at	108
	211377 x at			211372 s at	
	234376 at		S100A7	205916 at	254
	242026 at		S100A9	203535 at	255
NKX2-2	206915 at	249	CCND1	208711 s at	27
NKX2-8	207451 at	250		208712 at	
RAB34	1555630 a at	182		214019 at	
	224710 at		JAG2	209784 s at	115
MIF	217871 s at	134		32137 at	
GLI3	1569342 at	94	FOXM1	202580 x at	85
	205201 at		FOXF1	205935 at	83
	227376 at		FOXL1	216572 at	84
FST	204948 s at	87		243409 at	
	207345 at				
	226847 at				
BCL2	203684 s at	14			
	203685 at				
	207004 at				
	207005 s at				

Tabla 4. Lista curada con evidencia de los genes objetivo de la ruta AR utilizados en la red bayesiana y los conjuntos de sonda asociados usados para medir el nivel de expresión ARNm de los genes objetivo (# = número de secuencia en la lista de secuencias adjunta).

5

Genes objetivo	Conjunto de sondas	#	Genes objetivo	Conjunto de sondas	#
ABCC4	1554918 a at	2	LCP1	208885 at	128
	1555039 a at		LRIG1	211596 s at	132
	203196 at			238339 x at	
APP	200602 at	7	NDRG1	200632 s at	147
	211277 x at		NKX3-1	209706 at	251
	214953 s at			211497 x at	
AR	211110 s at	8		211498 s at	
	211621 at		NTS	206291 at	155
	226192 at		PLAU	205479 s at	167
	226197 at			211668 s at	
CDKN1A	1555186 at	34	PMEPA1	217875 s at	169
	202284 s at			222449 at	
CREB3L4	226455 at	42		222450 at	
DHCR24	200862 at	53	PPAP2A	209147 s at	171
DRG1	202810 at	58		210946 at	
EAF2	1568672 at	60	PRKACB	202741 at	175
	1568673 s at			202742 s at	
	219551 at			235780 at	
ELL2	214446 at	65	KLK3	204582 s at	123
	226099 at			204583 x at	
	226982 at		PTPN1	202716 at	180
FGF8	208449 s at	75		217686 at	
FKBP5	204560 at	77	SGK1	201739 at	195
	224840 at		TACC2	1570025 at	214
	224856 at			1570546 a at	
GUCY1A3	221942 s at	99		202289 s at	
	227235 at			211382 s at	
	229530 at		TMPRSS2	1570433 at	225
	239580 at			205102 at	
IGF1	209540 at	105		211689 s at	
	209541 at			226553 at	
	209542 x at		UGT2B15	207392 x at	236

ES 2 700 617 T3

	211577 s at			216687 x at	
KLK2	1555545 at	122			
	209854 s at				
	209855 s at				
	210339 s at				

Tabla 5. Símbolos genéticos de los genes objetivo ER que se encontraron con evidencia significativa en literatura (= lista de genes de objetivo ER) (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	#						
AP1B1	5	SOD1	205	MYC	142	ENSA	66
COX7A2L	41	TFF1	221	ABCA3	1	KIAA0182	118
CTSD	46	TRIM25	230	ZNF600	247	BRF1	19
DSCAM	59	XBP1	245	PDZK1	160	CASP8AP2	25
EBAG9	61	GREB1	97	LCN2	127	CCNH	29
ESR1	70	IGFBP4	106	TGFA	222	CSDE1	43
HSPB1	103	SGK3	196	CHEK1	38	SRSF1	213
KRT19	124	WISP2	241	BRCA1	18	CYP1B1	48
NDUFB3	148	ERBB2	69	PKIB	166	FOXA1	81
NRIP1	154	CA12	22	RET	188	TUBA1A	235
PGR	162	CELSR2	36	CALCR	23	GAPDH	91
PISD	164	CDH26	32	CARD10	24	SFI1	194
PRDM15	174	ATP5J	12	LRIG1	132	ESR2	258
PTMA	179	COL18A1	40	MYB	140	MYBL2	141
RARA	183	CCND1	27	RERG	187		

5

Tabla 6. Lista corta de los genes objetivo de Wnt con base en la puntuación de evidencia y relación de probabilidades en la literatura (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	#
KIAA1199	119
AXIN2	13
CD44	30
RNF43	189
MYC	142
TBX3	215
TDGF1	220
SOX9	209
ASCL2	10
IL8	110
SP5	210
ZNRF3	248
EPHB2	67
LGR5	131
EPHB3	68
KLF6	121
CCND1	27
DEFA6	52
FZD7	90

10 Tabla 7. Lista corta de los genes objetivo de ER con base en la puntuación de evidencia y relación de probabilidades en la literatura (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	#
CDH26	32
SGK3	196
PGR	162
GREB1	97
CA12	22
XBP1	244
CELSR2	36
WISP2	241
DSCAM	59
ERBB2	69
CTSD	46

TFF1	221
NRIP1	154

Tabla 8. Lista corta de los genes objetivo de Hedgehog con base en la puntuación de evidencia y relación de probabilidades en la literatura (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	#
GLI1	93
PTCH1	177
PTCH2	178
IGFBP6	107
SPP1	212
CCND2	28
FST	87
FOXL1	84
CFLAR	37
TSC22D1	232
RAB34	182
S100A9	255
S100A7	254
MYCN	144
FOXM1	85
GLI3	94
TCEA2	216
FYN	88
CTSL1	47

5 Tabla 9. Lista corta de los genes objetivo de AR con base en la puntuación de evidencia y relación de probabilidades en la literatura (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	#
KLK2	122
PMEPA1	169
TMPRSS2	225
NKX3-1	251
ABCC4	2
KLK3	123
FKBP5	77
ELL2	65
UGT2B15	236
DHCR24	53
PPAP2A	171
NDRG1	147
LRIG1	132
CREB3L4	42
LCP1	128
GUCY1A3	99
AR	8
EAF2	60

Ejemplo 4: Comparación de la lista curada de evidencia y la lista de literatura amplia

10 La lista de genes objetivo Wnt construidos con base en la evidencia de la literatura siguiendo el procedimiento descrito aquí (Tabla 1) se compara con otra lista de genes objetivo que no siguen el procedimiento mencionado anteriormente. La lista alternativa es una compilación de genes indicados por una variedad de datos de diversos enfoques experimentales para ser un gen Wnt objetivo publicado en tres fuentes públicas por reconocidos laboratorios, conocidos por su experiencia en el área de biología molecular y la ruta Wnt. La lista alternativa es una combinación de los genes mencionados en la tabla S3 de Hatzis et al. (Hatzis P, 2008), el texto y la tabla S1A de Sousa e Melo (de Sousa E Melo F, 2011) y la lista de genes objetivo recogidos y mantenidos por Roel Nusse, pionero en el campo de la señalización Wnt (Nusse, 2012). La combinación de estas tres fuentes dio lugar a una lista de 124 genes (=lista de literatura amplia, véase la Tabla 10). Aquí se discute la cuestión de si el rendimiento en la predicción de la actividad Wnt en muestras clínicas por el algoritmo derivado de esta lista alternativa está funcionando de manera similar o mejor en comparación con el modelo construido sobre la base de la lista existente de genes (=lista de curanderos probados, Tabla 1).

15

20

ES 2 700 617 T3

Tabla 10. Lista alternativa de genes objetivo de Wnt (= lista de literatura amplia) (# = número de secuencia en la lista de secuencias adjunta).

Genes objetivo	Referencia	#	Genes objetivo	Referencia	#
ADH6	de Sousa e Melo et al.	3	L1CAM	Nusse	125
ADRA2C	Hatzis et al.	4	LBH	Nusse	126
APCDD1	de Sousa e Melo et al.	6	LEF1	Hatzis et al., de Sousa e Melo et al., Nusse	130
ASB4	de Sousa e Melo et al.	9	LGR5	de Sousa e Melo et al., Nusse	131
ASCL2	Hatzis et al., de Sousa e Melo et al.	10	LOC283859	de Sousa e Melo et al.	260
ATOH1	Nusse	11	MET	Nusse	133
AXIN2	Hatzis et al., de Sousa e Melo et al., Nusse	13	MMP2	Nusse	135
BIRC5	Nusse	15	MMP26	Nusse	136
BMP4	Nusse	16	MMP7	Nusse	137
BMP7	Hatzis et al.	17	MMP9	Nusse	138
BTRC	Nusse	20	MRPS6	Hatzis et al.	139
BZRAP1	de Sousa e Melo et al.	21	MYC	Hatzis et al., Nusse	142
SBSPON	de Sousa e Melo et al.	259	MYCBP	Nusse	143
CCL24	de Sousa e Melo et al.	26	MYCN	Nusse	144
CCND1	Nusse	27	NANOG	Nusse	146
CD44	Nusse	30	NKD1	de Sousa e Melo et al.	150
CDH1	Nusse	31	NOS2	Nusse	151
CDK6	Hatzis et al.	33	NOTUM	de Sousa e Melo et al.	152
CDKN2A	Nusse	35	NRCAM	Nusse	153
CLDN1	Nusse	39	NUAK2	Hatzis et al.	156
COL18A1	Hatzis et al.	40	PDGFB	Hatzis et al.	159
CTLA4	Nusse	44	PFDN4	Hatzis et al.	161
CYP4X1	de Sousa e Melo et al.	49	PLAUR	Nusse	168
CYR61	Nusse	50	POU5F1	Nusse	170
DEFA5	de Sousa e Melo et al.	51	PPARD	Nusse	172
DEFA6	de Sousa e Melo et al.	52	PROX1	de Sousa e Melo et al.	176
DKK1	de Sousa e Melo et al., Nusse	54	PTPN1	Hatzis et al.	180
DKK4	de Sousa e Melo et al.	55	PTTG1	Nusse	181
DLL1	Nusse	56	REG3A	de Sousa e Melo et al.	185
DPEP1	de Sousa e Melo et al.	57	REG4	de Sousa e Melo et al.	186
EDN1	Nusse	62	RPS27	Hatzis et al.	190
EGFR	Nusse	64	RUNX2	Nusse	191
EPHB2	Hatzis et al., de Sousa e Melo et al., Nusse	67	SALL4	Nusse	192
EPHB3	Hatzis et al., Nusse	68	SLC1A1	de Sousa e Melo et al.	199
ETS2	Hatzis et al.	71	SLC7A5	Hatzis et al.	201
FAT1	Hatzis et al.	72	SNAI1	Nusse	202
FGF18	Nusse	73	SNAI2	Nusse	203
FGF20	Nusse	74	SNAI3	Nusse	204
FGF9	Nusse	76	SIK1	Hatzis et al.	261
FLAD1	Hatzis et al.	78	SOX17	Nusse	206
AK122582	Hatzis et al.	262	SOX2	de Sousa e Melo et al.	207
FN1	Nusse	79	SOX4	Hatzis et al.	208
FOSL1	Nusse	80	SOX9	Nusse	209
FOXP1	Nusse	86	SP5	Hatzis et al., de Sousa e Melo et al.,	210
FST	Nusse	87	SP8	Hatzis et al.	211
FZD2	de Sousa e Melo et al.	89	TCF3	Nusse	217
FZD7	Nusse	90	TDGF1	Hatzis et al.	220
GAST	Nusse	92	TIAM1	Nusse	224
GMDS	Hatzis et al.	96	TNFRSF19	Nusse	227
GREM2	Nusse	98	TNFSF11	Nusse	228
HES6	Hatzis et al.	100	TRIM29	de Sousa e Melo et al.	231
HNF1A	Nusse	102	TSPAN5	de Sousa e Melo et al.	233
ID2	Nusse	104	TTC9	de Sousa e Melo et al.	234
IL22	de Sousa e Melo et al.	109	VCAN	Nusse	237

ES 2 700 617 T3

IL8	Nusse	110	VEGFA	Nusse	238
IRX3	de Sousa e Melo et al.	111	VEGFB	Nusse	239
IRX5	de Sousa e Melo et al.	112	VEGFC	Nusse	240
ISL1	Nusse	113	WNT10A	Hatzis et al.	242
JAG1	Nusse	114	WNT3A	Nusse	243
JUN	Nusse	116	ZBTB7C	de Sousa e Melo et al.	246
KIAA1199	de Sousa e Melo et al.	119	PATZ1	Hatzis et al.	263
KLF4	Hatzis et al.	120	ZNRF3	Hatzis et al.	248

El siguiente paso consistió en encontrar los conjuntos de sonda de la matriz Affymetrix® GeneChip Human Genome U133 Plus 2.0 que corresponden con los genes. Este proceso se realizó utilizando el complemento Bioconductor en R y curado manual en cuanto a la relevancia de conjuntos de sonda con base en el navegador de genoma de UCSC, eliminando por lo tanto, por ejemplo, conjuntos de sonda en las cadenas opuestas o fuera de las regiones de exón del gen. Para dos de los 124 genes no hay conjuntos de sonda disponibles en este microarreglo chip y por lo tanto no podrían ser insertados en la red Bayesiana, siendo estos LOC283859 y WNT3A. Se encontró que en total 287 conjuntos de sonda corresponden a los restantes 122 genes (Tabla 11).

- 5
- 10 Tabla 11. Conjuntos de sonda asociados con los genes objetivo Wnt en la lista genética amplia de la literatura (# = número de secuencia en la lista de secuencias adjunta).

Símbolo genético	Conjunto de sonda	#	Símbolo genético	Conjunto de sonda	#	Símbolo genético	Conjunto de sonda	#
ADH6	207544 s at	3	FAT1	201579 at	72	PFDN4	205360 at	161
	214261 s at		FGF18	206987 x at	73		205361 s at	
ADRA2C	206128 at	4		211029 x at			205362 s at	
APCDD1	225016 at	6		211485 s at		PLAUR	210845 s at	168
ASB4	208481 at	9		231382 at			211924 s at	
	217228 s at		FGF20	220394 at	74		214866 at	
	217229 at		FGF9	206404 at	76	POU5F1	208286 x at	170
	235619 at			239178 at		PPARD	208044 s at	172
	237720 at		FLAD1	205661 s at	78		210636 at	
	237721 s at			212541 at			37152 at	
ASCL2	207607 at	10	AK122582	235085 at	262		242218 at	
	229215 at		FN1	1558199 at	79	PROX1	207401 at	176
ATOH1	221336 at	11		210495 x at			228656 at	
AXIN2	222695 s at	13		211719 x at		PTPN1	202716 at	180
	222696 at			212464 s at			217686 at	
	224176 s at			214701 s at			217689 at	
	224498 x at			214702 at		PTTG1	203554 x at	181
BIRC5	202094 at	15		216442 x at		REG3A	205815 at	185
	202095 s at		FOSL1	204420 at	80		234280 at	
	210334 x at		FOXN1	207683 at	86	REG4	1554436 a at	186
BMP4	211518 s at	16	FST	204948 s at	87		223447 at	
BMP7	209590 at	17		207345 at		RPS27	200741 s at	190
	209591 s at			226847 at		RUNX2	216994 s at	191
	211259 s at		FZD2	210220 at	89		221282 x at	
	211260 at			238129 s at			232231 at	
BTRC	1563620 at	20	FZD7	203705 s at	90		236858 s at	
	204901 at			203706 s at			236859 at	
	216091 s at		GAST	208138 at	92	SALL4	229661 at	192
	222374 at		GMDS	204875 s at	96	SLC1A1	206396 at	199
	224471 s at			214106 s at			213664 at	
BZRAP1	205839 s at	21	GREM2	220794 at	98	SLC7A5	201195 s at	201
SBSPON	214725 at	259		235504 at		SNAI1	219480 at	202
	235209 at			240509 s at		SNAI2	213139 at	203
	235210 s at		HES6	226446 at	100	SNAI3	1560228 at	204
CCL24	221463 at	26		228169 s at		SIK1	208078 s at	261
CCND1	208711 s at	27	HNF1A	210515 at	102		232470 at	
	208712 at			216930 at		SOX17	219993 at	206
	214019 at		ID2	201565 s at	104		230943 at	
CD44	1557905 s at	30		201566 x at		SOX2	213721 at	207
	204489 s at			213931 at			213722 at	
	204490 s at		IL22	221165 s at	109		228038 at	
	209835 x at			222974 at		SOX4	201416 at	208

ES 2 700 617 T3

	210916 s at		IL8	202859 x at	110		201417 at	
	212014 x at			211506 s at			201418 s at	
	212063 at		IRX3	229638 at	111		213668 s at	
	217523 at		IRX5	210239 at	112	SOX9	202935 s at	209
	229221 at		ISL1	206104 at	113		202936 s at	
CDH1	201130 s at	31	JAG1	209097 s at	114	SP5	235845 at	210
	201131 s at			209098 s at		SP8	237449 at	211
	208834 x at			209099 x at			239743 at	
CDK6	207143 at	33		216268 s at		TCF3	209151 x at	217
	214160 at		JUN	201464 x at	116		209152 s at	
	224847 at			201465 s at			209153 s at	
	224848 at			201466 s at			210776 x at	
	224851 at		KIAA1199	1554685 a at	119		213730 x at	
	231198 at			212942 s at			213811 x at	
	235287 at		KLF4	220266 s at	120		215260 s at	
	243000 at			221841 s at			216645 at	
CDKN2A	207039 at	35	L1CAM	204584 at	125	TDGF1	206286 s at	220
	209644 x at			204585 s at		TIAM1	206409 at	224
	211156 at		LBH	221011 s at	126		213135 at	
CLDN1	218182 s at	39	LEF1	210948 s at	130	TNFRSF19	223827 at	227
	222549 at			221557 s at			224090 s at	
COL18A1	209081 s at	40		221558 s at		TNFSF11	210643 at	228
	209082 s at		LGR5	210393 at	131		211153 s at	
CTLA4	221331 x at	44		213880 at		TRIM29	202504 at	231
	231794 at		MET	203510 at	133		211001 at	
	234362 s at			211599 x at			211002 s at	
	236341 at			213807 x at		TSPAN5	209890 at	233
CYP4X1	227702 at	49		213816 s at			213968 at	
CYR61	201289 at	50	MMP2	1566678 at	135		225387 at	
	210764 s at			201069 at			225388 at	
DEFA5	207529 at	51	MMP26	220541 at	136	TTC9	213172 at	234
DEFA6	207814 at	52	MMP7	204259 at	137		213174 at	
DKK1	204602 at	54	MMP9	203936 s at	138	VCAN	204619 s at	237
DKK4	206619 at	55	MRPS6	224919 at	139		204620 s at	
DLL1	224215 s at	56	MYC	202431 s at	142		211571 s at	
	227938 s at		MYCBP	203359 s at	143		215646 s at	
DPEP1	205983 at	57		203360 s at			221731 x at	
EDN1	218995 s at	62		203361 s at		VEGFA	210512 s at	238
	222802 at		MYCN	209756 s at	144		210513 s at	
EGFR	1565483 at	64		209757 s at			211527 x at	
	1565484 x at			211377 x at			212171 x at	
	201983 s at			234376 at		VEGFB	203683 s at	239
	201984 s at		NANOG	220184 at	146	VEGFC	209946 at	240
	210984 x at		NKD1	1553115 at	150	WNT10A	223709 s at	242
	211550 at			229481 at			229154 at	
	211551 at			232203 at		ZBTB7C	217675 at	246
	211607 x at		NOS2	210037 s at	151	ZBTB7C	227782 at	246
EPHB2	209588 at	67	NOTUM	228649 at	152	PATZ1	209431 s at	263
	209589 s at		NRCAM	204105 s at	153		211391 s at	
	210651 s at			216959 x at			210581 x at	
	211165 x at		NUAK2	220987 s at	156		209494 s at	
EPHB3	1438 at	68	PDGFB	204200 s at	159	ZNRF3	226360 at	248
	204600 at			216061 x at				
ETS2	201328 at	71		217112 at				
	201329 s at							

Posteriormente, la red Bayesiana se construyó de forma similar a la Figura 6 y el procedimiento se explica en el presente documento. De forma similar a la descripción de la red bayesiana de Wnt con base en la lista de curado de evidencia, las tablas de probabilidad condicional de los bordes entre conjuntos de sonda y sus genes respectivos, tanto la lista de curado de evidencia como la lista de literatura general, fueron entrenadas usando datos procesados por fRMA de 32 muestras de colon normales y 32 muestras de adenoma del conjunto de datos GSE8671 de la Gene Expression Omnibus (accesible en <http://www.ncbi.nlm.nih.gov/geo/>, visitado por última vez el 13 de julio de 2011).

5

Las redes Bayesianas entrenadas fueron probadas en varios conjuntos de datos para inferir la probabilidad P (Wnt Habilitada) de que la ruta Wnt está "habilitada", es decir, activa, la cual se toma igual a la probabilidad inferida de que el complejo de transcripción de la ruta Wnt está "presente". En las Figuras 14-19 se muestran los resultados resumidos del modelo de literatura amplia y el modelo curado de evidencia.

Evidentemente, podría deducirse que el modelo de literatura general predice generalmente probabilidades más extremas para que la señalización de Wnt sea activada o desactivada. Además, el modelo alternativo predice resultados similares para los conjuntos de datos de cáncer de colon (GSE20916, GSE4183, GSE15960), pero muestras más que esperadas con conjuntos de datos de predicción de la señalización Wnt activa en cáncer de mama (GSE12777), cáncer de hígado (GSE9843) y muestra de meduloblastoma (GSE10327).

En conclusión, la lista amplia de la literatura de genes da como resultado predicciones aproximadamente igual de buenas de actividad Wnt en el cáncer de colon, por un lado, pero peores predicciones (demasiados falsos positivos) en otros tipos de cáncer, por otro lado. Esto podría ser un resultado de que la lista alternativa de los genes objetivo está demasiado sesgada hacia células de colon específicamente, por lo tanto, demasiado específicas para tejidos; el interés principal, tanto de Sousa E Melo et al. como de Hatzis et al. era el cáncer colorrectal aunque se pueden incluir genes de Wnt no específicos de colon. Además, los genes objetivo no específicos de Wnt posiblemente incluidos en estas listas pueden ser una fuente de las predicciones empeoradas de la actividad de Wnt en otros tipos de cáncer. Es probable que la lista alternativa contenga genes objetivo más regulados indirectamente, lo que probablemente la haga más específica para tejidos. La lista original está ajustada para contener genes objetivo directos, los cuales son más propensos a representar genes que son sensibles a Wnt en todos los tejidos, reduciendo así la especificidad para tejidos.

#### Ejemplo 5: Entrenamiento y uso de la red Bayesiana

Antes de que se pueda utilizar la red Bayesiana para inferir la actividad de la ruta en una muestra de ensayo, se han de determinar los parámetros que describen las relaciones probabilísticas entre los elementos de la red. Además, en el caso de estados discretos de las mediciones de entrada, deben establecerse umbrales que describen cómo realizar la discretización.

Típicamente, las redes Bayesianas son entrenadas usando un conjunto representativo de muestras de entrenamiento, de las cuales se conocen preferentemente todos los estados de todos los nodos de la red. Sin embargo, no es práctico obtener muestras de entrenamiento de muchos tipos diferentes de cánceres, de los cuales se sabe cuál es el estado de activación de la ruta que va a ser modelada. Como resultado, los conjuntos de entrenamiento disponibles consisten en un número limitado de muestras, típicamente de un tipo de cáncer solamente. Para permitir que la red Bayesiana se genere bien a otros tipos de muestras, se debe prestar especial atención a la forma en que se determinan los parámetros, lo que se realiza preferiblemente de la siguiente manera en el enfoque descrito en el presente documento.

Para el nodo TF, la probabilidad (incondicional) de estar en estado "ausente" y "presente" es dada por la aparición esperada en un gran conjunto de muestras. Alternativamente, se puede establecer a 0.5, como se hace en la Figura 1, con el fin de no tener sesgo para un resultado positivo o negativo.

Para los nodos de genes objetivo, las probabilidades condicionales se establecen como en la Figura 1. Si el elemento TF está "ausente", es más probable que el gen objetivo esté "abajo", por lo tanto se elige una probabilidad de 0.95 para este y probabilidad de 0.05 para el gen objetivo que está "arriba". Esta última probabilidad (no nula) es para tener en cuenta la (rara) posibilidad de que el gen objetivo sea regulado por otros factores u observado accidentalmente "hacia arriba" (por ejemplo debido al ruido de medición). Si el elemento TF está "presente", entonces con una probabilidad justa de 0.70 el gen objetivo está "arriba", y con una probabilidad de 0.30 el gen objetivo está "abajo". Estos últimos valores se eligen de esta manera, porque puede haber varias razones por las que un gen objetivo no esté muy expresado aunque el elemento TF esté presente, por ejemplo porque la región promotora del gen está metilada. En el caso de que un gen objetivo no esté regulado por el elemento TF, pero regulado a la baja, las probabilidades se eligen de manera similar, pero reflejan la regulación a la baja ante la presencia del elemento TF.

Para el modelo de red Bayesiana, tal como se muestra en la Figura 6, donde las intensidades de los conjuntos de sonda forman las mediciones de entrada, finalmente se tienen que determinar los parámetros para la discretización y para las tablas de probabilidad condicional que relacionan las intensidades de los conjuntos de sonda con los niveles de ARNm de Los respectivos genes objetivo. Ambos se basan en datos de entrenamiento en la presente invención. Para la discretización del nivel de intensidad de un conjunto de sondas en los estados "bajo" y "alto", se determina un umbral adecuado que mejor separa los valores de intensidad en un conjunto de muestras de entrenamiento donde se activa la ruta (muestras "habilitadas") de los valores de intensidad en un conjunto de muestras de entrenamiento en el que no está. Muestras deshabilitadas). Por último, las tablas de probabilidades condicionales que describen las probabilidades de un conjunto de sondas para tener una intensidad "baja" o "alta" dependiendo del estado "bajo" o "alto" del gen objetivo respectivo, se realizan contando el número de muestras "habilitadas" y "deshabilitadas" con un valor de intensidad del conjunto de sondas por debajo y por encima del umbral respectivo. Esto se conoce en la literatura como el enfoque de frecuencias. Se agrega un conteo ficticio a cada grupo para evitar entradas en las tablas de probabilidad condicional con un valor de cero, para evitar un comportamiento extremo de la red Bayesiana.

Después de que la red Bayesiana ha sido entrenada, puede aplicarse en una muestra de prueba de la siguiente manera, considerando la red Bayesiana de la Figura 6, y asumiendo que las mediciones por microarreglos relacionadas con los conjuntos de sonda están disponibles. El primer paso es discretizar las mediciones de entrada, comparando la intensidad de cada conjunto de sondas en la muestra de prueba con el umbral respectivo como se ha descrito anteriormente. Esta comparación puede hacerse de una manera difícil, poniendo cada conjunto de sondas a una intensidad "baja" o "alta" (llamada "evidencia dura"), o puede hacerse de una manera suave, asumiendo cierta incertidumbre (ruido) en la medición, estableciendo para cada conjunto de sondas una probabilidad de ser "baja" o "alta" (llamada "evidencia suave"). Por ejemplo, la prueba suave de un conjunto de sondas con una intensidad justo debajo del umbral puede ser una probabilidad de 0.8 de ser "baja" y una probabilidad de 0.2 de ser "alta", con base en una estimación adecuada del ruido y la diferencia con respecto al umbral.

A continuación, esta evidencia dura o suave se suministra a un motor de inferencia adecuado para redes Bayesianas, por ejemplo, con base en un algoritmo de árbol de cruce (véase (Napolitano, 2004)). Dicho motor puede deducir la probabilidad actualizada de que el elemento TF esté "ausente" o "presente", dada la evidencia proporcionada. La probabilidad deducida de que el elemento TF esté "presente" se interpreta entonces como la probabilidad estimada de que la ruta respectiva esté activa.

Preferiblemente, el entrenamiento de los modelos de red Bayesiana de las rutas Wnt, ER, Hedgehog y AR se realiza utilizando datos públicos disponibles en el Gene Expression Omnibus (accesible en <http://www.ncbi.nlm.nih.gov/geo/>, ver arriba).

La red Bayesiana Wnt fue entrenada como ejemplo utilizando 32 muestras normales de colon consideradas como una ruta Wnt inactiva y 32 muestras confirmadas de adenoma que se sabe que tienen una ruta Wnt activa (conjunto de datos GSE8671).

El modelo de red Bayesiana de la ruta ER fue entrenado como ejemplo usando 4 muestras de MCF7 desprovistas de estrógenos, que se sabe que tienen una ruta ER inactiva, y 4 muestras de MCF7 estimuladas con estrógenos, las que se considera que tienen una ruta ER activa, a partir del conjunto de datos de GSE8597 también accesible en la Gene Expression Omnibus.

El modelo de red Bayesiana de la ruta Hedgehog fue entrenado como ejemplo usando 15 muestras de carcinoma de células basales de las que se confirmó que tenían una ruta Hedgehog activa y 4 muestras de células de piel normales representando muestras con una ruta Hedgehog inactiva disponible en el conjunto de datos de GSE7553.

El modelo de la red Bayesiana de la ruta AR fue entrenado como ejemplo usando 3 muestras con actividad AR positiva, líneas celulares LNCaP estimuladas con dihidrotestosterona (DHT), un potente activador de la ruta AR y 3 líneas celulares LNCaP no estimuladas que representan la AR inactiva caso de ruta.

Con referencia a la Figura 35 y la Figura 36, los modelos de la red Bayesiana entrenados de la ruta Wnt y ER fueron utilizados para predecir las actividades de la ruta en muestras similares (muestras de colon y MCF7 para la red Wnt y ER Bayesiana respectivamente) no utilizados en el procedimiento de entrenamiento como se describe aquí (no se encontró ningún conjunto de datos apropiado para la red Bayesiana de Hedgehog). Las actividades previstas de la ruta de la gran mayoría de las muestras deben estar en línea con las actividades clínicamente esperadas de la ruta para que el modelo sea validado.

La Figura 35 muestra las actividades Wnt previstas, representadas como el logit de P (Wnt habilitada) sobre el eje vertical, para las muestras, ilustradas por las barras en el eje horizontal, de las muestras de colon agrupadas por clasificación, indicadas por el color de la barra, en el conjunto de datos GSE20916. Se ha predicho correctamente que todas las muestras normales de colon tienen una ruta inactiva (puntuación <0), sobre la base de que es una muestra de tejido sano. Se prevé que todas las muestras, excepto cuatro, que tienen una ruta activa, tienen una ruta Wnt activa.

En la Figura 36 se muestran los resultados de validación del modelo de red Bayesiana ER ensayado para dos microarreglos medidos utilizando una línea de células de cáncer de mama MCF7, una estimulada con estrógeno (E2) y la otra con un control negativo (EtOH), procedente del conjunto de datos GSE9253. De acuerdo con la supuesta actividad de ER, se prevé que la muestra estimulada con estrógeno tenga una ruta ER activa, mientras que el control negativo predice una ruta ER inactiva.

Más detalles y ejemplos para usar redes Bayesianas entrenadas (por ejemplo, Wnt, ER, AR y ruta Hedgehog) para predecir las respectivas actividades de la ruta se explican en el Ejemplo 6 a continuación.

El procedimiento de formación mencionado anteriormente puede emplearse para otras redes Bayesianas de aplicaciones clínicas. Aquí está demostrado y probado que funciona para los modelos de red Bayesiana construidos utilizando el método descrito en el presente documento que representa rutas de señalización celular, más específicamente las rutas Wnt, ER, AR y Hedgehog.

Ejemplo 6: Diagnóstico de la actividad de la ruta (anormal)

A continuación se ilustrará a manera de ejemplo cómo usar por ejemplo modelos de red Bayesiana para diagnosticar la actividad de una ruta de señalización celular.

5 Las redes Bayesianas de la ruta Wnt, ER, Hedgehog y AR, construidas utilizando un nodo para la presencia del factor de transcripción, una capa de nodos que representan el ARNm de los genes objetivo y una capa de nodos que representan las intensidades de los conjuntos de sonda correspondientes a los genes objetivo (Tabla 1, Tabla 2, Tabla 3 y Tabla 4), análogos a la Figura 6 descrita aquí, y entrenados como se describe en este documento, fueron utilizadas para predecir la actividad de las rutas como "habilitadas", esto es, activas, o "inhabilitadas", esto es, inactivas, en conjuntos de datos diversos, previamente no utilizados para entrenamiento, para inferir cómo funciona bien el componente de inferencia. Los puntajes de la actividad de la ruta prevista se correlacionan con el conocimiento clínico. Los resúmenes de los resultados para una selección de los ensayos se muestran en las Figuras 21 sec.

15 Con referencia a las Figuras 21 sec., se muestran resultados de inferencia de actividad de trayectoria para muestras de tejido médico usando el modelo de red Bayesiana descrito aquí.

La Figura 21 muestra los resultados de las pruebas de actividad de Wnt en el conjunto de datos de muestras de colon GSE4183. El modelo de la red Bayesiana arrojó valores altos de P (Wnt habilitada) para las muestras de adenoma y valores bajos para muestras normales, lo que corresponde con la (patofisiología del adenoma y el tejido sano. El tejido sano tiene una proliferación celular lenta y por lo tanto una actividad baja de Wnt con relación al tejido adenomatoso que tiene una proliferación celular rápida y por lo tanto una alta actividad de Wnt. Para las muestras IBD, el modelo de la red Bayesiana mostró baja actividad de la ruta Wnt (P (Wnt habilitada) ~0) para todas las muestras excepto una. De nuevo, esto es consistente con las muestras de IBD que no experimentan proliferación celular rápida. Para las muestras de células de cáncer colorrectal los resultados se mezclaron, con alta actividad de la ruta Wnt que se detecta en aproximadamente la mitad de estas muestras, pero esto puede ser el resultado de otras rutas que asumen el papel de conductoras del tumor cuando el tejido adenomatoso benigno se convierte en tejido canceroso maligno, o análisis de muestras, por ejemplo, conteniendo la muestra demasiado tejido no tumoral, o estando parcialmente degradado el ARNm.

El modelo de red Bayesiana utilizado en los experimentos aquí informados se formó utilizando el conjunto de datos de muestras de colon GSE8671. Sin embargo, la ruta Wnt está presente (aunque posiblemente inactiva) en otros tipos de células. Por lo tanto, se consideró posible que la red Bayesiana podría ser aplicable para inferir anormalmente alta actividad de la ruta Wnt correlativa con otros tipos de cánceres. La razón de esto es que, aunque el modelo de la red Bayesiana fue entrenado usando muestras de colon, se basa en los primeros principios de la operación de la ruta Wnt presente (aunque posiblemente inactiva) en otros tipos de células. Las Figuras 22-24 muestran algunos resultados que investigan tales inferencias "tipo tejido cruzado".

La Figura 22 muestra los resultados de las pruebas utilizando el modelo de red Bayesiana entrenado usando muestras de colon que se aplican para inferir la actividad de la ruta Wnt en muestras de meduloblastoma (conjunto de datos GSE10327). Las muestras incluidas en este conjunto de datos se han caracterizado adicionalmente en varios subconjuntos, siendo uno de ellos muestras con la ruta Wnt activa. La red bayesiana de la ruta Wnt predice el grupo de Wnt activo muestras que tienen una ruta Wnt activa, mientras que se predijo correctamente que las otras muestras tenían una ruta Wnt inactiva.

En la Figura 23 se muestran los resultados de las pruebas utilizando el modelo de red Bayesiana Wnt en un conjunto de datos que contiene muestras de cáncer de hígado (GSE9843). Aquí las muestras se agrupan mediante las siguientes anotaciones a priori asignadas por el conjunto de datos GSE9843: "CTNNB1", "Inflamación", "Polisomía chr7", "Proliferación" y "Sin anotación". Se infieren uniformemente que las muestras del grupo de "Inflamación" no tienen una actividad anormalmente alta de la ruta Wnt, como se esperaba ya que la condición de inflamación no implica proliferación celular rápida. También se infiere uniformemente que las muestras etiquetadas "Polisomía chr7" no tienen una actividad de ruta Wnt anormalmente alta. La polisomía del cromosoma número 7 significa que hay más de dos cromosomas número 7. Como no hay razón para esperar esta condición de polisomía para impactar la ruta Wnt, no es inesperado que estas muestras no tengan actividad de ruta Wnt anormalmente alta.

Aproximadamente una de cada cinco de las muestras marcadas con "Proliferación" tiene P (Wnt habilitada) >0,5. La proliferación sugiere un estado de rápida multiplicación celular. Tal estado puede estar asociado con una actividad anormalmente alta de la ruta Wnt, pero también puede estar asociado con numerosas otras posibles causas de proliferación celular. Por consiguiente, aproximadamente una de cada cinco de estas muestras que tienen una actividad de ruta Wnt anormalmente alta no es un resultado irrazonable.

Aproximadamente para la mitad de las muestras del grupo "CTNNB 1" la red Bayesiana infiere que tienen una actividad de la ruta Wnt anormalmente alta. El gen CTNNB 1 codifica la proteína beta-catenina, que es una proteína reguladora de la ruta Wnt, y las mutaciones activadoras en este gen causan la activación anormal de Wnt. Por lo tanto, una correlación entre el grupo "CTNNB 1" y actividad alta de la ruta Wnt es la expectativa de conformidad.

La Figura 24 representa los resultados de ensayo del modelo de red Bayesiana de Wnt descrito aquí para un conjunto de muestras de cáncer de mama. En este caso, se ensayan tres grupos de líneas celulares de cáncer de mama: un grupo para el cual se conoce a priori que la ruta Wnt funciona a un nivel anormalmente alto (grupo con Wnt habilitada); un grupo para el cual la ruta Wnt es a priori conocida por no estar operando a un nivel anormalmente alto (grupo con Wnt deshabilitada); y otro grupo para el cual la actividad de la ruta Wnt no se conoce a priori (grupo desconocido); además hay una muestra que se sospecha que tiene un bajo nivel de activación Wnt (Wnt sospechoso), aunque hay un informe contradictorio en la literatura de que puede tener una ruta Wnt activa (pero este es un informe minoritario, más artículos reportan una ruta Wnt inactiva). Como se ve en la Figura 24, la correlación de las inferencias proporcionadas por la red Bayesiana con el conocimiento a priori es fuerte para los grupos de Wnt habilitada y deshabilitada. También la muestra más a la derecha del gráfico (Wnt sospechoso) muestra una inferencia que corresponde a la mayoría de los informes en la literatura que indican que la ruta Wnt está inhabilitada. En el caso del grupo desconocido que se muestra en la Figura 24, para el cual no hay conocimiento a priori de la actividad de la ruta Wnt, la red Bayesiana infiere baja actividad para la ruta Wnt excepto para una instancia para la cual  $P(\text{Wnt habilitada}) > 0,5$ ; la literatura muestra que esta línea celular tiene una alta expresión del correceptor LRP6, lo que puede explicar que la ruta Wnt esté habilitada.

La Figura 25 muestra los resultados para el mismo conjunto de datos de líneas celulares de cáncer de mama, pero ahora se ensaya la actividad ER utilizando la red Bayesiana ER entrenada usando líneas celulares de cáncer de mama MCF7 como se describe en el presente documento. Se predijo que las muestras a priori conocidas por tener una ruta de Wnt activa tenían una ruta ER inactiva, lo cual no es sorprendente ya que la ruta de Wnt ya está impulsando la multiplicación rápida de células. Las muestras positivas para ER, por otro lado se encuentran entre las muestras con Wnt deshabilitadas y las muestras desconocidas. En vista de la Figura 24, esto no es sorprendente.

En la Figura 26 se muestran los resultados de las predicciones de la red Bayesiana ER entrenada en líneas celulares de cáncer de mama para un conjunto de muestras de cáncer (GSE12276). Las muestras de cáncer de mama se subdividieron en las clasificaciones bien conocidas: Luminal A (LumA), Luminal B (LumB), receptor del factor de crecimiento epidérmico humano 2 positivo (HER2) y subtipo basal de cáncer de mama. Se sabe que las muestras de tejido en los subtipos luminal A y luminal B expresan ER. Es también en estos subtipos que se prevé que la mayoría de las muestras tengan una actividad de ruta ER alta. Por otro lado, se sabe que las muestras que se clasifican como del subtipo basal tienen una expresión baja o inexistente de ER, lo que se correlaciona bien con una ruta ER no activa predicha en las muestras del grupo basal. En el grupo HER2 solo tres muestras tienen un  $P(\text{ER habilitada}) > 0,5$ , mientras que se prevé que la mayoría de las muestras tienen una ruta ER inactiva. Esto se correlaciona bien con el hecho de que la clasificación se hace sobre el hecho de que estas muestras tienen una expresión amplificada de HER2; la replicación celular no controlada es presumiblemente conducida a través de la señalización de HER2 a través de otras rutas de señalización celular diferentes a la ruta ER (véanse por ejemplo las líneas celulares de cáncer de mama activas Wnt en la Figura 24 o las muestras activas de cáncer de mama Hedgehog en la Figura 30).

El modelo de red Bayesiana ER construido y entrenado como se describe en el presente documento se utiliza para predecir la actividad de la ruta ER en un gran panel de líneas celulares de diversos cánceres, cuyos resultados se muestran en la Figura 27. Como era de esperar, las muestras con predicción de ER se encontraron en las líneas de células de cáncer de mama. Se predijo que todos los otros tipos de líneas celulares de cáncer tenían una ruta ER inactiva, que es como se esperaba.

El modelo de la red Bayesiana construido y entrenado para la ruta Hedgehog como se describe aquí se utiliza para predecir la actividad de la ruta Hedgehog para las líneas celulares de diversos tipos de cáncer en el conjunto de datos GSE34211. Las predicciones de actividad de Hedgehog se muestran en la Figura 28. Las más altas fracciones de la actividad de Hedgehog pronosticada se encuentran en los tipos de cáncer del sistema nervioso central (SNC), piel, endometrio y útero, lo que está de acuerdo con el conocimiento de la literatura con respecto a la proliferación celular dependiente de Hedgehog en estos tipos de células.

La Figura 29 muestra la actividad de Hedgehog pronosticada de las muestras de meduloblastoma (GSE10327) que ya se analizó usando el modelo de red Bayesiana de Wnt como se describe en el presente documento. Las muestras de meduloblastoma se han caracterizado en subclases, teniendo una de ellas una ruta de señalización activa de Hedgehog (identificador: SHH). Se prevé que todas las muestras en el subtipo SHH tengan una señalización activa de Hedgehog. Además, se predijo que las muestras de meduloblastoma en el subtipo Wnt también tenían una ruta Hedgehog activa. Esto está de acuerdo con la evidencia clínica que demuestra que a menudo ambas rutas son activas en estos tumores. Sin embargo, la red Bayesiana de Wnt fue claramente capaz de predecir correctamente la actividad Wnt solo en el subtipo Wnt. Así, la combinación de la red Bayesiana Wnt y Hedgehog es capaz de hacer una clasificación correcta de estos dos subtipos.

En la Figura 30 se muestra la predicción de la actividad de Hedgehog en las muestras de cáncer de mama GSE12276, utilizada anteriormente para predecir la actividad de ER usando el modelo de red Bayesiana ER, utilizando el modelo de red Bayesiana Hedgehog. Se prevé que la ruta Hedgehog esté activa en una fracción de las muestras de cada subtipo. Esto parece extraño, pero en concordancia con la predicción de la ruta ER mostrada en la Figura 26, se puede ver que la actividad de Hedgehog solo se predice en muestras que no tienen una ruta ER activa. Esto está en buen

acuerdo con la hipótesis de que la proliferación celular no controlada en el tejido (de mama) puede ser impulsada por diferentes rutas de señalización.

5 En resumen, los resultados de las pruebas para varias muestras de tejido canceroso y células presentadas en las Figuras 21-30 sugieren fuertemente que las redes Bayesianas de los modelos Wnt, ER y Hedgehog entrenadas en muestras específicas de tejido/ruta son aplicables al análisis de muestras de otros tipos de tejido. Esto puede permitir que el análisis de la ruta de señalización celular se aplique al "tipo de tejido cruzado". Por lo tanto, el sistema CDS 10 (como se describe en el presente documento) se aplica fácilmente para evaluar la actividad de la ruta en un intervalo de tipos de tejidos distintos del tipo de tejido de las muestras utilizadas para formar el modelo de red Bayesiana 40 (véase, por ejemplo, (CDS) configurado para evaluar una o más rutas de señalización celular como se describe aquí (mostrado como ejemplo para la ruta Wnt)). En los casos en que los componentes de inferencia 40, 44, 46, 48 indiquen que el tejido analizado exhibe una actividad de rutas Wnt, ER o Hedgehog anormalmente alta, pero no hay fármaco específico de tejido disponible, el médico puede considerar un fármaco general de supresión de rutas Wnt, ER o Hedgehog con base en la recomendación 28 o en la recomendación 26, respectivamente, según lo dispuesto por el sistema CDS 10.

Aunque los resultados de las Figuras 21-30 indican la aplicabilidad del tipo de tejido cruzado del modelo de red Bayesiana para las rutas Wnt, ER y Hedgehog, se espera que para aplicaciones clínicas los modelos de red Bayesiana se puedan actualizar o adaptar opcionalmente para maximizar su aplicabilidad al tipo de tejido específico bajo análisis (por ejemplo, tejido mamario o tejido hepático). Tal actualización o adaptación podría, por ejemplo, implicar el ajuste de las probabilidades condicionales con base en estudios clínicos del tipo de tejido bajo análisis o enriquecer la lista de genes objetivo curada con evidencia, descrita en el presente documento, con genes objetivo específicos de tejido de la ruta investigada. Adicionalmente, pueden eliminarse o agregarse nodos para ajustar mejor el modelo de la red Bayesiana al tejido bajo análisis. Alternativamente, pueden entrenarse diferentes modelos de redes Bayesianas ab 20 inicio usando diferentes conjuntos de entrenamiento para los diferentes tipos de tejidos. Además, los resultados de las Figuras 21-30 ilustran la capacidad del procedimiento descrito en el presente documento para desarrollar y entrenar modelos de red Bayesiana usando listas de genes objetivo curadas con evidencia de rutas diferentes a Wnt, ER y Hedgehog para predecir y diagnosticar la actividad de la ruta.

30 Los resultados de ensayo del modelo de red Bayesiana AR construido y entrenado como se describe aquí se usaron a manera de ejemplo para predecir la actividad de AR en líneas celulares de cáncer de próstata LNCaP tratadas con diferentes regímenes de tratamiento (GSE7708) (véase la Figura 38). Como se esperaba, las células LNCaP no estimuladas con DHT resultan en una ruta AR predecible inactiva, mientras que se predijo correctamente que las células estimuladas con LNCaP tenían una ruta AR activa y las células LNCaP tratadas con Bicalutamida, un fármaco antiandrogénico, tenían una ruta AR inhibida.

La red Bayesiana entrenada de la ruta AR como se describe aquí también se usó para predecir la probabilidad de que la ruta AR sea activa en muestras de cáncer de próstata del conjunto de datos GSE17951 (los resultados se muestran en la Figura 39). Inesperadamente no se predijo que la mayoría de las biopsias y los tumores de próstata tenían una mayor probabilidad de actividad AR en comparación con las muestras de los controles.

45 El modelo de la red Bayesiana AR también se aplicó a una prueba de tejidos cruzados, a saber, las muestras de cáncer de mama incluidas en el conjunto de datos GSE12276. Los resultados de esta prueba se muestran en la Figura 40. Se prevé que una pequeña fracción de las muestras, que se encuentran en cada subgrupo, tiene una ruta activa, mientras que la gran mayoría de las muestras tenían una ruta AR inactiva. Sorprendentemente, el porcentaje más alto de muestras con una ruta AR activa se encuentra en el subgrupo HER2, lo cual no es inesperado ya que se sabe de la literatura que hay interferencia entre la ruta HER2 y AR y la ruta AR también puede ser inducida por la señalización de HER2-.

50 El modelo de red Bayesiana AR antes mencionado también se usó para predecir la actividad de la ruta AR en dos conjuntos de muestras de líneas celulares de diversos tipos de cáncer (GSE36133 y GSE34211), como se representa en la Figura 41 y Figura 42. Como era de esperar, se encontró que la mayoría de las líneas celulares tenían una ruta AR inactiva. Las excepciones a esto son las muestras de cáncer de próstata con varias muestras de líneas celulares de cáncer que expresan la actividad de la ruta AR. En la Tabla 12 se muestra que todas las predicciones de la actividad de la ruta AR de las muestras de cáncer de próstata están de acuerdo con la actividad AR conocida.

Tabla 12. Actividad AR conocida y predicha en líneas celulares de cáncer de próstata en conjuntos de datos GSE36133 y GSE34211.

Conjunto de datos	Identificador de la muestra	Línea celular de próstata	¿Se sabe que está activo?	P (AR habilitada)
36133	GSM886837	22Rv1	SI	0.698127
	GSM886988	DU 145	NO	0.001279
	GSM887271	Clon LNCaP FGC	SI	1

	GSM887302	MDA PCa 2b	SI	1
	GSM887440	NCI-H660	NO	1.25E-05
	GSM887506	PC-3	NO	0.009829
	GSM887731	VCaP	SI	1
34211	GSM843494	DU145T	NO	0.005278
	GSM844559	HPET11	NO	0.005602
	GSM844560	HPET13 replicado 1	NO	0.003382
	GSM844561	HPET13 replicado 2	NO	0.000501
	GSM844562	HPET5	NO	0.007673
	GSM844579	LNCAP	SI	1
	GSM844674	PC3 PFIZER	NO	0.004066
	GSM844675	PC3 Good_NCI50_WYETH	NO	0.006163

Ejemplo 7: Pronóstico con base en la actividad de la ruta

5 Se cree que las rutas de desarrollo temprano, como Wnt y Hedgehog, juegan un papel en la metástasis causada por las células cancerosas que han revertido a un fenotipo más similar a células madre, llamado células madre de cáncer. De hecho, existen pruebas suficientes de que las rutas de desarrollo precoces, como la ruta Wnt, juegan un papel en las metástasis del cáncer, permitiendo que las células cancerosas metastásicas comiencen a dividirse en la localización de siembra en otro órgano o tejido. La metástasis está asociada con un mal pronóstico, por lo que se espera que la actividad de las rutas de desarrollo precoces, como la ruta Wnt y Hedgehog, en las células cancerosas sea predictiva del mal pronóstico. Esto es apoyado por el hecho de que los pacientes con cáncer de mama, del conjunto de datos GSE12276, que fueron identificados con una ruta ER activa, pero que no tienen una ruta Wnt o Hedgehog activa utilizando los modelos de red Bayesiana descritos en este documento, tenía un mejor pronóstico que los pacientes identificados que tienen una ruta Hedgehog Wnt o ambas activa, como se ilustra en la gráfica de Kaplan-Meier de la Figura 37.

Ejemplo 8: Planificación de la terapia, predicción de la eficacia del fármaco, predicción de efectos adversos y monitorización de la eficacia del fármaco

20 El ejemplo siguiente ilustra cómo utilizar los modelos probabilísticos, en particular los modelos de red Bayesiana, para la planificación de la terapia, la predicción de la eficacia del fármaco, la supervisión de la eficacia del fármaco y las actividades relacionadas.

25 El modelo de la red bayesiana de la ruta ER, construido utilizando un nodo para la presencia del factor de transcripción, una capa de nodos que representan los niveles de ARNm de los genes objetivo (Tabla 2) y una capa de nodos que representan las intensidades de los sondas correspondientes a los genes objetivo (Tabla 2), análogo a la Figura 6 descrita en este documento, y entrenado como se describe en el presente documento, se usaron para predecir la actividad de la ruta. Posteriormente, se demuestra que la actividad de la ruta está correlacionada con la eficacia del fármaco o la eficacia del fármaco de control. Los resúmenes de los resultados se muestran en las Figuras 31 y 32.

30 El tamoxifeno es un fármaco utilizado actualmente para el tratamiento del cáncer de mama ER+ (receptor de estrógeno positivo). Actúa como un antagonista parcial del receptor de estrógenos que inhibe la proliferación celular incontrolada que se cree que es inducida por la señalización ER. Desafortunadamente, no todos los cánceres de mama responden al tratamiento con tamoxifeno, a pesar de la demostración de la presencia de proteína ER en células cancerosas mediante el análisis histopatológico de rutina de secciones de tejido de cáncer. Se han realizado muchos estudios para investigar esta llamada resistencia al tamoxifeno. El conjunto de datos GSE21618 disponible públicamente es el resultado de uno de dichos estudios y contiene datos de microarreglos de líneas de células MCF7 resistentes a tamoxifeno y de tipo salvaje bajo diferentes regímenes de tratamiento. El modelo de red Bayesiana ER construido y entrenado como se describe aquí se usa para analizar las líneas celulares resistentes a tamoxifeno y MCF7 bajo diferentes regímenes de tratamiento, cuyos resultados se muestran en la Figura 31.

45 Se prevé que la línea celular resistente a tamoxifeno, indicada por TamR.Ctrl, tenga una ruta ER inactiva para cada punto de tiempo después de la adición de tamoxifeno (1, 2, 3, 6, 12, 24 y 48 h). No es de extrañar que el tratamiento de la línea celular resistente al tamoxifeno, que es insensible al tratamiento con Tamoxifeno, con Tamoxifeno, indicado por TamR.Tam, es ineficaz, lo cual también se ilustra por la inactividad prevista de la ruta ER para este grupo sobre los mismos puntos de tiempo. De acuerdo con el análisis de la línea de células resistentes al tamoxifeno (TamR.Ctrl), la fuerza motriz de la proliferación celular no controlada no se debe a la señalización ER activa; por lo tanto, el tratamiento con un antagonista de ER no inhibe la proliferación celular. Esto ilustra que el tratamiento con tamoxifeno no se recomienda en el caso de una actividad de ruta ER predecible.

50 Por otra parte, la línea celular MCF7 de tipo salvaje, conocida como sensible al tamoxifeno, tratada con 17beta-estradiol (wt1.E2) reacciona lentamente al tratamiento hormonal que es visible en las predicciones crecientes de actividad positiva de ER. El tratamiento de dicha línea celular con inhibidores de la aromatasa que se sabe que inhiben la producción de estrógenos, inhibirá la ruta ER, que se ilustra por la disminución de la predicción de la ruta ER en el

tiempo. Apoyan esto las predicciones de ruta ER hechas con base en los datos de microarreglos de muestras MCF7 tratadas con estrógeno para aumentar el tiempo en el conjunto de datos GSE11324, mostrándose los resultados en la Figura 32.

5 Lo anterior ilustra la capacidad de los modelos probabilísticos, en particular los modelos de red Bayesiana, para ser utilizados en la planificación de la terapia, la predicción de la eficacia del fármaco y la monitorización de la eficacia del fármaco. Sin embargo, debe entenderse que la misma metodología también se aplicaría para predecir y controlar los efectos adversos.

10 Ejemplo 9: Desarrollo de fármacos

Similar a la monitorización de la respuesta a la terapia, puede utilizarse un modelo de ruta en el desarrollo de fármacos para evaluar la eficacia de diversos compuestos putativos. Por ejemplo, al escanear muchos compuestos para un posible efecto sobre una cierta ruta en una línea celular de cáncer, el modelo de ruta respectiva puede usarse para determinar si la actividad de la ruta sube o baja después de la aplicación del compuesto o no. A menudo, esta comprobación se realiza utilizando solo uno o algunos de los marcadores putativos de la actividad de la ruta, lo que aumenta la probabilidad de un seguimiento ineficaz del efecto del tratamiento. Además, en los estudios de seguimiento en animales o sujetos pacientes, los modelos de ruta pueden usarse de manera similar para evaluar la eficacia de los fármacos candidatos y para determinar una dosis óptima para afectar al máximo la actividad de la ruta.

Un ejemplo de monitorización ineficaz de nuevos compuestos de fármaco se ilustra mediante la actividad de la ruta AR predicha en las muestras de GSE7708 como se muestra en la Figura 38. En este estudio se han desarrollado dos posibles compuestos de fármacos para inhibir la actividad de la ruta AR, denominados Poliamida 1 y Poliamida 2. Se ha demostrado que estas dos poliamidas son capaces de inhibir la ruta AR con base en los hallazgos de que las poliamidas se unen al elemento de respuesta a los andrógenos (ARE) e inhiben la expresión de KLK3 (= PSA), un marcador bien conocido para la actividad AR también incluidos en la selección de genes objetivo tal como se describe en este documento, así como ≈35% de los transcritos que fueron inducidos por DHT. En contraste, el modelo de la red Bayesiana de la ruta AR predijo que estas muestras todavía tienen una ruta AR activa. La investigación de las probabilidades inferidas de los genes objetivo que se sobreregularan utilizando el modelo de red Bayesiana para AR indicó que KLK3 en contraste con los otros genes objetivo fue subregulado de acuerdo con los hallazgos, mientras que todos los otros genes objetivo (excepto AR, GUCY1A3 y TMPRSS2 en el caso de la poliamida 1) se expresaron claramente diferencialmente en las muestras tratadas con poliamida 1 y poliamida 2. En otras palabras, solo un marcador para la actividad de AR, KLK3, fue subregulado, mientras que la mayoría de los genes objetivo identificados todavía se sobreregularon indicando que la ruta AR sigue estando en gran medida intacta y, por tanto, activa. Teniendo en cuenta un mayor número de genes objetivo con base en la evidencia bibliográfica, los inventores fueron capaces de demostrar que la inhibición de la actividad AR de las poliamidas es limitada y que solo la expresión KLK3 está claramente subregulada usando estas poliamidas. Por otra parte, esto ilustra el valor de un enfoque sistemático utilizando un modelo de red Bayesiana en comparación con un enfoque reduccionista en el desarrollo de fármacos.

40 Ejemplo 10: Desarrollo de ensayo

En lugar de aplicar las redes Bayesianas mencionadas en los datos de entrada de ARNm procedentes de microarreglos o secuenciación de ARN, puede ser beneficioso en aplicaciones clínicas desarrollar ensayos dedicados para realizar las mediciones de muestra, por ejemplo en una plataforma integrada usando qPCR para determinar los niveles de ARNm de genes objetivo. Las secuencias de ARN/ADN de los genes objetivo descritos pueden usarse entonces para determinar qué cebadores y sondas se seleccionan en dicha plataforma.

La validación de tal ensayo dedicado se puede hacer usando las redes Bayesianas con base en microarreglos como un modelo de referencia y verificando si el ensayo desarrollado da resultados similares en un conjunto de muestras de validación. Junto a un ensayo dedicado, esto también se puede hacer para construir y calibrar modelos de red Bayesiana similares usando datos de secuenciación de ARNm como medidas de entrada.

Ejemplo 11: Investigación de la ruta y investigación de la fisiopatología del cáncer

55 A continuación se ilustra cómo pueden emplearse modelos de redes Bayesianas en la investigación (clínica) de rutas, esto es, la investigación interesada en descubrir qué rutas están implicadas en ciertas enfermedades, las cuales pueden ser seguidas para una investigación más detallada, por ejemplo para vincular las mutaciones en las proteínas de señalización a los cambios en la activación de la ruta (medida con el modelo). Esto es relevante para investigar la iniciación, el crecimiento y la evolución y la metástasis de los cánceres específicos (la fisiopatología).

60 Los modelos de la red Bayesiana de las rutas Wnt, ER, Hedgehog y AR, construidos utilizando un nodo para la presencia del factor de transcripción, una capa de nodos que representan los niveles de ARNm de los genes objetivo (Tabla 1, Tabla 2, Tabla 3 y Tabla 4) y una capa de nodos que representan las intensidades de las sondas que corresponden a los genes objetivo (Tabla 1, Tabla 2, Tabla 3 y Tabla 4), análogas a la Figura 6 descritas aquí y entrenadas como se describe en el presente documento, se usó para predecir la actividad de la ruta de un conjunto de datos consistente en muestras de cáncer de mama (GSE12276).

Supongamos que el investigador está interesado en investigar la ruta o caminos de señalización celular y la(s) desregulación(es) específica(s) que impulsa(n) la proliferación celular incontrolada. El investigador puede analizar los datos de microarreglos utilizando los modelos probabilísticos antes mencionados, en particular los modelos de red Bayesiana, para encontrar qué rutas son presumiblemente la causa de la proliferación celular incontrolada. Se muestra en la Figura 33 y en la Figura 34 una ilustración de dicho análisis para el caso de la actividad de Wnt, ER y Hedgehog (muestras basales y luminal A del conjunto de datos GSE12276). Posteriormente, el investigador puede buscar en más detalle para encontrar la causa exacta de la desregulación de la ruta.

Con referencia a la Figura 34, se sabe que las muestras basales tienen un estado de receptor triple negativo (ER, PR y HER2), por lo tanto no es sorprendente ver que se predice que todas las muestras tienen una ruta ER inactiva. Por otra parte, se prevé que algunas de las muestras tienen tanto Wnt como Hedgehog o ambas activas como se muestra en la Figura 34. Estas actividades de la ruta predicha persuaden al investigador a investigar estas muestras con más detalle para, por ejemplo, mutaciones conocidas u otras desregulaciones conocidas en las rutas Wnt y/o Hedgehog.

Otro ejemplo se da en la Figura 33, en la que se ilustran las actividades Wnt, ER y Hedgehog en las muestras luminal A del conjunto de datos GSE12276. Se sabe que las muestras de Luminal A expresan ER, sin embargo esto no significa necesariamente que las propiedades cancerosas se deban a señalización ER activa. De las actividades de la ruta prevista se puede inferir que menos de la mitad de las muestras tienen una señalización ER activa. Sin embargo, se encuentran que algunas de las muestras que no tienen una señalización ER activa tienen una ruta Wnt y/o Hedgehog activa. Esto podría dar lugar a que el investigador explore estas muestras en detalles más cercanos en cuanto a los defectos en la ruta de señalización Wnt y/o Hedgehog, respectivamente. Algunas de las muestras no predicen ninguna de las tres rutas incluidas que son activas; tal vez otros caminos están causando las proliferaciones no controladas de células. También esto da al investigador información adicional para buscar defectos en otras rutas.

En resumen, las ilustraciones aquí descritas indican la capacidad de modelos de red Bayesiana entrenados (como se ha descrito anteriormente) para apoyar el proceso de encontrar la causa de la proliferación celular no controlada en un método más dirigido. Mediante el empleo de las redes Bayesianas para cribar las muestras con respecto a las actividades de la ruta, las actividades de la ruta prevista pueden identificar las rutas posibles para la proliferación celular, que pueden ser seguidas para una investigación más detallada, por ejemplo para vincular las mutaciones en las proteínas de señalización u otras desregulaciones conocidas a los cambios en la activación (medida con el modelo).

Tal como se describe en el presente documento, el proceso para desarrollar y formar una red Bayesiana de rutas de señalización celular puede usarse para construir un modelo de red Bayesiana para otras rutas que también podrían emplearse en relación con la presente invención.

Ejemplo 12: Inscripción del sujeto en un ensayo clínico con base en la actividad prevista

Si un fármaco candidato se desarrolla para, por ejemplo, bloquear la actividad de una determinada ruta que impulsa el crecimiento tumoral, y este fármaco está entrando en ensayos clínicos, resulta entonces esencial una selección adecuada de los sujetos que van a ser inscritos en tal ensayo es esencial para probar la efectividad potencial del fármaco. En tal caso, los pacientes que no tienen la ruta respectiva activada en sus tumores deben ser excluidos del ensayo, ya que es obvio que el fármaco no puede ser eficaz si la ruta no se activa en primer lugar. Por lo tanto, un modelo de ruta que puede predecir la actividad de la ruta se puede utilizar como una herramienta de selección, para seleccionar solo los pacientes que se prevé que tienen activada la ruta respectiva.

Ejemplo 13: Selección de las pruebas posteriores por realizar

Si se analiza un tumor utilizando diferentes modelos de rutas y los modelos predicen la desregulación de una cierta ruta, lo cual puede guiar entonces la selección de las pruebas posteriores por realizar. Por ejemplo, se puede ejecutar un ensayo de ligación de proximidad (PLA) para confirmar la presencia del respectivo complejo de transcripción (Söderberg O, 2006). Tal PLA puede diseñarse para dar un resultado positivo si dos proteínas clave en un complejo TF se han unido, por ejemplo, betacatenina y TCF4 en el complejo TF de la ruta Wnt.

Otro ejemplo es que la ruta cuya desregulación se ha predicho se analiza con más detalle con respecto a la cascada de señalización. Por ejemplo, se pueden analizar proteínas clave en esta ruta para determinar si hay mutaciones en las regiones de ADN que codifican para sus respectivos genes, o se puede probar la abundancia de estas proteínas para ver si son más alta o más baja de lo normal. Tales pruebas pueden indicar cuál es la causa raíz detrás de la desregulación de la ruta y dar una idea de los fármacos disponibles que podrían utilizarse para reducir la actividad de la ruta.

Estos ensayos se seleccionan para confirmar la actividad de la ruta identificada utilizando el modelo Bayesiano. Sin embargo, también es posible la selección de pruebas de diagnóstico complementarias. Después de la identificación de la ruta utilizando el modelo, para la elección de terapia dirigida solo deben realizarse las pruebas de diagnóstico complementario (la selección), que son aplicables a la ruta identificada.

## Ejemplo 14: Selección de pruebas de diagnóstico asociadas

De manera similar al ejemplo anterior, si se analiza un tumor y los modelos de ruta predicen la desregulación de una determinada ruta, y opcionalmente se han realizado una serie de pruebas adicionales para investigar la causa de la desregulación, entonces un oncólogo puede seleccionar un número de fármacos candidatos para tratar al paciente. Sin embargo, el tratamiento con dicho fármaco puede requerir una prueba diagnóstica complementaria que se debe ejecutar primero, por ejemplo, para cumplir con las pautas clínicas o para asegurar el reembolso de los costes del tratamiento, o porque se requiere regulación (FDA) para llevar a cabo la prueba de diagnóstico acompañante para suministrar el fármaco. Un ejemplo de esta prueba de diagnóstico complementaria es la prueba Her2 para el tratamiento de pacientes con cáncer de mama con el fármaco Herceptin (Trastuzumab). Por lo tanto, el resultado de los modelos de ruta se puede utilizar para seleccionar los fármacos candidatos y las respectivas pruebas de diagnóstico complementarias que se han de realizar.

## Ejemplo 15: Aplicación del CDS

Haciendo referencia a la Figura 20 (que muestra esquemáticamente un sistema de soporte de decisión clínica (CDS) configurado para evaluar una o más rutas de señalización celular como se describe aquí (mostrado como ejemplo para la ruta Wnt)), se implementa un sistema 10 de soporte de decisión clínica (CDS) como un ordenador 12 adecuadamente configurado. El ordenador 12 puede estar configurado para funcionar como el sistema CDS 10 ejecutando un software, un firmware u otras instrucciones adecuados almacenados en un medio de almacenamiento no transitorio (no mostrado) tal como un disco duro u otro medio de almacenamiento magnético, un disco óptico u otro medio de almacenamiento óptico, una memoria de acceso aleatorio (RAM), memoria de solo lectura (ROM), memoria instantánea u otro medio de almacenamiento electrónico, un servidor de red, etc. Aunque el sistema 10 CDS ilustrativo es ejecutado por el ordenador 12 ilustrativo, más generalmente el sistema CDS puede ser ejecutado por un dispositivo de procesamiento digital o un aparato que comprende un procesador digital configurado para llevar a cabo métodos de soporte de decisión clínica como se expone aquí. Por ejemplo, el dispositivo de procesamiento digital puede ser un dispositivo portátil (por ejemplo, un asistente de datos personales o un teléfono inteligente que ejecuta una aplicación CDS), un ordenador portátil, un ordenador de escritorio, un ordenador o dispositivo de tableta, un servidor de red remoto, etc. El ordenador 12 u otro dispositivo de procesamiento digital típicamente incluye o está conectado operativamente con un dispositivo 14 de visualización a través del cual se muestra información que incluye recomendaciones de soporte de decisión clínica al personal médico. El ordenador 12 u otro dispositivo de procesamiento digital típicamente también incluye o está conectado operativamente con uno o más dispositivos de entrada de usuario, tal como un teclado 16ilustrativo, o un ratón, bola de desplazamiento, alfombrilla de desplazamiento, pantalla táctil (posiblemente integrada con el dispositivo 14 de visualización), u otro dispositivo de entrada de usuario con base en puntero, a través del cual el personal médico puede introducir información tal como órdenes operacionales para controlar el sistema CDS 10, datos para su uso por el sistema CDS 10, etc.

El sistema CDS 10 recibe como información de entrada perteneciente a un sujeto médico (por ejemplo, un paciente del hospital o un paciente ambulatorio que está siendo tratado por un oncólogo, médico u otro personal médico o una persona sometida a cribado de cáncer o algún otro diagnóstico médico conocido, o de quien se sospecha que tiene cierto tipo de cáncer, como cáncer de colon, cáncer de mama o cáncer de hígado, etc.). El sistema CDS 10 aplica diversos algoritmos de análisis de datos a esta información de entrada para generar recomendaciones de soporte de decisión clínica que son presentadas al personal médico a través del dispositivo 14 de visualización (o a través de un sintetizador de voz u otro dispositivo que proporcione una salida perceptible por el usuario). En algunas realizaciones, estos algoritmos pueden incluir la aplicación de una guía clínica al paciente. Una guía clínica es un conjunto almacenado de recomendaciones de tratamiento estándar o "canónicas", construidas típicamente sobre la base de recomendaciones de un panel de expertos médicos y opcionalmente formateadas en forma de un "diagrama de flujo" clínico para facilitar la navegación a través de la guía clínica. En diversas realizaciones, los algoritmos de procesamiento de datos del CDS 10 pueden incluir adicionalmente o alternativamente diversos algoritmos de pruebas de diagnóstico o clínicas que se realizan sobre la información de entrada para extraer recomendaciones de decisión clínica, tales como métodos de aprendizaje de máquinas descritos en el presente documento.

En los sistemas CDS ilustrativos descritos en el presente documento (por ejemplo, el sistema CDS 10), los algoritmos de análisis de datos de CDS incluyen uno o más algoritmos de diagnóstico o de prueba clínica que se ejecutan con la información genómica y/o proteómica de entrada adquirida por uno o más laboratorios 18 médicos. Los laboratorios pueden ubicarse de forma diversa "en el lugar", es decir, en el hospital u otro lugar donde el sujeto médico está siendo sometido a un examen y/o tratamiento médico, o "fuera del sitio", por ejemplo un laboratorio especializado y centralizado que recibe (por correo u otro servicio de entrega) una muestra de tejido del sujeto médico que se ha extraído del sujeto médico (por ejemplo, una muestra obtenida de una lesión de mama o de un colon de un sujeto médico de quien se sabe o sospecha que tiene cáncer de colon, o de un hígado de un sujeto médico de quien se sabe o sospecha que tiene cáncer de hígado, etc., mediante un procedimiento de biopsia u otro procedimiento de extracción de muestras). El tejido del que se extrae una muestra puede ser también tejido metastásico, por ejemplo tejido maligno (sospechado) procedente del colon, mama, hígado u otro órgano, que se ha diseminado fuera del colon, la mama, el hígado u otro órgano. En algunos casos, la muestra de tejido puede ser células tumorales circulantes, es decir, células tumorales que han entrado en el torrente sanguíneo y pueden extraerse como la muestra de tejido extraída usando técnicas de aislamiento adecuadas. La muestra extraída es procesada por el laboratorio para generar información

genómica o proteómica. Por ejemplo, la muestra extraída puede ser procesada usando un microarreglo (también denominado en la técnica como un chip genético, chip de ADN, biochip, etc.) o mediante un procesamiento cuantitativo de reacción en cadena de la polimerasa (qPCR) para medir la información genómica o proteómica probatoria tal como los niveles de expresión de los genes de interés, por ejemplo en forma de un nivel de ácido ribonucleico mensajero (ARNm) que se transcribe a partir del gen, o un nivel de una proteína que se traduce del ARNm transcrito del gen. Como otro ejemplo, la muestra extraída puede ser procesada por un laboratorio de secuenciación génica para generar secuencias de ácido desoxirribonucleico (ADN), o para generar una secuencia de ARN, variación de número de copias, etc. Otros enfoques de medición contemplados incluyen inmunohistoquímica (IHC), citología, hibridación *in situ* por fluorescencia (FISH), ensayo de ligación de proximidad, etc., realizados en una sección patológica. Otra información que puede ser generada por el procesamiento de microarreglos, espectrometría de masas, secuenciación de genes u otras técnicas de laboratorio incluye información de metilación. También se pueden realizar diversas combinaciones de tales mediciones genómicas y/o proteómicas.

En algunas realizaciones, los laboratorios 18 médicos realizan una serie de adquisiciones de datos estandarizadas sobre la muestra extraída del tejido del sujeto médico, con el fin de generar una gran cantidad de datos genómicos y/o proteómicos. Por ejemplo, las técnicas de adquisición de datos estandarizadas pueden generar una secuencia de ADN (opcionalmente alineada) para uno o más cromosomas o porciones cromosómicas, o para todo el genoma del tejido. La aplicación de un microarreglo estándar puede generar miles o decenas de miles de elementos de datos, como niveles de expresión para un gran número de genes, varios datos de metilación, y así sucesivamente. Esta plétora de datos genómicos y/o proteómicos, o partes seleccionadas de los mismos, se introducen en el sistema CDS 10 para ser procesados de manera que se desarrolle información útil desde el punto de vista clínico para formular recomendaciones de soporte de decisión clínica.

Los sistemas CDS descritos y los métodos relacionados se refieren al procesamiento de datos genómicos y/o proteómicos para evaluar la actividad de diversas rutas de señalización celular. Sin embargo, debe entenderse que los sistemas CDS descritos (por ejemplo, el sistema CDS 10) opcionalmente pueden incluir adicionalmente diversas capacidades adicionales, tales como generar recomendaciones de soporte de decisión clínica de acuerdo con guías clínicas almacenadas con base en diversos datos de pacientes tales como monitorización de signos vitales, datos del paciente, datos demográficos del paciente (por ejemplo, sexo, edad, etc.), datos de imágenes médicas del paciente, etc. Alternativamente, en algunas realizaciones las capacidades del sistema CDS 10 pueden limitarse solamente a realizar análisis de datos genómicos y/o proteómicos para evaluar rutas de señalización celular como se describe en el presente documento.

Continuando con referencia a la Figura 20 de ejemplo, el sistema CDS 10 infiere la actividad de una ruta de señalización celular en el tejido del sujeto médico con base al menos en, pero sin limitarse a, niveles de expresión de genes objetivo de la ruta de señalización celular medida en la muestra y determina si la ruta de señalización celular está funcionando anormalmente en el tejido del sujeto médico con base en esta actividad inferida. Los ejemplos descritos en el presente documento se refieren a las rutas Wnt, ER, AR y Hedgehog como rutas de señalización celular ilustrativas. Estas rutas son de interés en diversas áreas de la oncología porque la pérdida de la regulación de las rutas puede ser una causa de la proliferación de un cáncer. Hay aproximadamente 10-15 rutas de señalización relevantes, y cada cáncer es impulsado por, en principio, una ruta dominante que está siendo desregulada. Sin limitarse a ninguna teoría particular de la operación, estas rutas regulan la proliferación celular y, consecuentemente, una pérdida de regulación de estas rutas en las células cancerosas puede conducir a que la ruta esté "siempre habilitada", acelerando así la proliferación de células cancerosas, que a su vez se manifiesta como un crecimiento, invasión o metástasis (propagación) del cáncer.

La medición de los niveles de expresión de ARNm de genes que codifican para proteínas reguladoras de la ruta de señalización celular, tal como una proteína intermedia que forma parte de una cascada de proteínas que forma la ruta de señalización celular, es una medida indirecta del nivel de expresión de la proteína reguladora y puede o puede no correlacionarse fuertemente con el nivel real de expresión de la proteína reguladora (mucho menos con la actividad global de la ruta de señalización celular). La ruta de señalización celular regula directamente la transcripción de los genes objetivo -por lo tanto, los niveles de expresión de ARNm transcritos a partir de los genes objetivo son un resultado directo de esta actividad reguladora-. Por lo tanto, el sistema CDS 10 infiere la actividad de la ruta de señalización celular (por ejemplo, las rutas Wnt, ER, AR y Hedgehog) con base al menos en los niveles de expresión de genes objetivo (nivel de ARNm o proteína como medida sustituta) de la ruta de señalización celular. Esto asegura que el sistema CDS 10 infiere la actividad de la ruta con base en información directa proporcionada por los niveles de expresión medidos a partir de los genes objetivo.

Sin embargo, si bien, como se describe en el presente documento, son eficaces para evaluar la actividad de las rutas globales, los niveles 20 de expresión medidos de los genes objetivo de las rutas no son especialmente informativos en cuanto a por qué las rutas están funcionando anormalmente (si es así). Dicho de otro modo, los niveles 20 de expresión medidos de los genes objetivo de una ruta pueden indicar que la ruta está funcionando anormalmente, pero no indican qué parte de la ruta está funcionando mal (por ejemplo, carece de regulación suficiente) para hacer que la ruta general funcione anormalmente.

Por consiguiente, si el sistema CDS 10 detecta una actividad anormal de una ruta particular, el sistema CDS 10 hace uso opcionalmente de otra información proporcionada por los laboratorios 18 médicos para la muestra extraída, tal como secuencias 22 genéticas alineadas y/o medidas(s) de expresión para uno o más genes reguladores de la ruta 24, o seleccionar la prueba diagnóstica que se va a realizar a continuación para evaluar qué parte de la ruta está funcionando mal. Para maximizar la eficacia, en algunas realizaciones esta evaluación opcional de por qué la ruta está funcionando incorrectamente se lleva a cabo solamente si el análisis de los niveles 20 de expresión medidos de los genes objetivo de la ruta indica que la ruta está funcionando anormalmente. En otras realizaciones, esta evaluación se integra en el análisis probabilístico de la ruta de señalización celular descrita en el presente documento.

En las realizaciones en las que el sistema CDS 10 evalúa qué parte de la ruta está funcionando mal y tiene éxito al hacerlo, la información adicional permite al sistema CDS 10 recomendar la prescripción de un fármaco dirigido al mal funcionamiento específico (recomendación 26 mostrada en la Figura 20). Si no se identifica ningún mal funcionamiento específico de la ruta (ya sea porque la evaluación adicional opcional no se realiza o porque esa evaluación no identifica ninguna porción particular de la ruta que está funcionando mal), entonces el sistema CDS 10 puede proporcionar una recomendación 28 por defecto recomendando la prescripción de un fármaco de supresión general para esta ruta particular (suponiendo que la actividad de la ruta anormal es actividad excesivamente alta).

Ejemplo 16: Un kit y herramientas de análisis para medir la actividad de la ruta

El conjunto de genes objetivo que han mostrado una mejor indicación de la actividad de la ruta específica, con base en la investigación con base en la secuenciación de microarreglos/ARN utilizando el modelo Bayesiano, puede traducirse en un ensayo de PCR cuantitativo multiplex para ser realizado en una muestra de tejido o célula. Para desarrollar una prueba aprobada por la FDA para la actividad de la ruta, se requiere el desarrollo de un kit de prueba estandarizado, que debe ser validado clínicamente en ensayos clínicos para obtener la aprobación regulatoria.

En general, debe entenderse que mientras que los ejemplos que son pertinentes a las rutas Wnt, ER, AR y Hedgehog se proporcionan como ejemplos ilustrativos, los enfoques para el análisis de la ruta de señalización celular descritos aquí se aplican fácilmente a otras rutas de señalización celular además de estas rutas, tales como las rutas de señalización intracelular con receptores en la membrana celular (ver más arriba) y las rutas de señalización intracelular con receptores dentro de la célula (ver más arriba). Además: Esta solicitud describe varias realizaciones preferidas. Pueden colegirse por otras modificaciones y alteraciones al leer y entender la descripción detallada precedente. Se pretende que la solicitud se interprete con la inclusión de todas las dichas modificaciones y alteraciones en la medida en que entren dentro del alcance de las reivindicaciones adjuntas o sus equivalentes.

Literatura:

de Sousa E Melo F, C. S. (2011). Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. *Cell Stem Cell.*, 476-485

Hatzis P, v. d. (2008). Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol.*, 2732-2744

Neapolitan, R. (2004). *Learning Bayesian networks*. Pearson Prentice Hall

Nusse, R. (2012, May 1). Wnt target genes. Retrieved from The Wnt homepage: [http://www.stanford.edu/group/nusselab/cgi-bin/wnt/target\\_genes](http://www.stanford.edu/group/nusselab/cgi-bin/wnt/target_genes)

Söderberg O, G. M. (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. *Nat Methods.*, 995-1000

van de Wetering M, S. E.-P.-F. (2002). The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell*, 241-250.

**REIVINDICACIONES**

1. Un método que comprende:

- 5 inferir la actividad de una o más rutas de señalización celular en tejido de un sujeto médico con base al menos en el nivel o niveles (20) de expresión de uno o más genes objetivo de las rutas de señalización celular medidos en una muestra extraída del tejido del sujeto médico, en donde la inferencia comprende:
- 10 inferir la actividad de las rutas de señalización celular en el tejido del sujeto médico evaluando al menos una porción de un modelo probabilístico (40-1, ..., 40-7), preferiblemente una red Bayesiana (40-1, ..., 40-7), que representan las rutas de señalización celular para un conjunto de entradas que incluyen al menos los niveles (20) de expresión del uno o más genes objetivo de las rutas de señalización celular medidas en la muestras extraída del tejido del sujeto médico;
- 15 estimar un nivel (46) en el tejido del sujeto médico de al menos un elemento de factor de transcripción (TF), el al menos un elemento TF que controla la transcripción del uno o más genes objetivo de las rutas de señalización celular, estando basada la estimación al menos en parte en probabilidades condicionales que relacionan el al menos un elemento TF y los niveles (20) de expresión de uno o más genes objetivo de las rutas de señalización celular medidas en la muestra extraída del tejido del sujeto médico; e
- 20 inferir la actividad de la ruta de señalización celular con base en el nivel estimado en la muestra de tejido del elemento de factor de transcripción;
- 25 en el que la inferencia se realiza mediante un dispositivo (12) de procesamiento digital que utiliza el modelo probabilístico (40-1, ..., 40-7) de las rutas de señalización celular,

2. Un método de acuerdo con la reivindicación 1, en el que la inferencia comprende:

- 30 estimar un nivel (46) en el tejido del sujeto médico de al menos un elemento de factor de transcripción (TF) representado por un nodo TF del modelo probabilístico, controlando el elemento TF la transcripción del uno o más genes objetivo de las rutas de señalización celular, estando basada la estimación al menos en parte en las probabilidades condicionales del modelo probabilístico (40-1, ..., 40-7) relacionando el nodo TF y los nodos en el modelo probabilístico que representa el uno o más genes objetivo de las rutas de señalización celular, medidos en la muestra extraída del tejido del sujeto médico,
- 35 y en el que la inferencia se realiza preferiblemente utilizando una red Bayesiana (40-1, ..., 40-7) que comprende nodos que representan información sobre la ruta de señalización y relaciones de probabilidad condicional entre nodos conectados de la red Bayesiana.

40 3. El método de una cualquiera de las reivindicaciones 1 o 2, en el que las rutas de señalización celular comprenden una ruta Wnt, una ruta ER, una ruta AR y/o una ruta Hedgehog.

4. El método de la reivindicación 3, en el que la inferencia comprende:

- 45 inferir la actividad de una ruta Wnt en el tejido del sujeto médico con base en al menos en los niveles (20) de expresión de uno o más, preferiblemente al menos tres, genes objetivos de la ruta Wnt medida en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende: KIAA1199, AXIN2, RNF43, TBX3, TDGF1, SOX9, ASCL2, IL8, SP5, ZNRF3, KLF6, CCND1, DEFA6 y FZD7.

50 5. El método de la reivindicación 3 o 4, en el que la inferencia (también) comprende:

- inferir la actividad de una ruta ER en el tejido del sujeto médico con base en al menos en los niveles (20) de expresión de uno o más, preferiblemente al menos tres, genes objetivo de la ruta ER medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende: CDH26, SGK3, PGR, GREB1, CA12, XBP1, CELSR2, WISP2, DSCAM, ERBB2, CTSD, TFF1 y NRIP1.
- 55

6. El método de cualquiera de las reivindicaciones 3-5, en el que la inferencia (también) comprende:

- 60 inferir la actividad de una ruta Hedgehog en el tejido del sujeto médico con base en al menos en los niveles (20) de expresión de uno o más, preferiblemente al menos tres, genes objetivo de la ruta Hedgehog medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende: GLI1, PTCH1, PTCH2, IGFBP6, SPP1, CCND2, FST, FOXL1, CFLAR, TSC22D1, RAB34, S100A9, S100A7, MYCN, FOXM1, GLI3, TCEA2, FYN y CTSL1.

7. El método de cualquiera de las reivindicaciones 3-6, en el que la inferencia (también) comprende:

inferir la actividad de una ruta AR en el tejido del sujeto médico con base en al menos en los niveles (20) de expresión de uno o más, preferiblemente al menos tres, genes objetivo de la ruta AR medidos en la muestra extraída del tejido del sujeto médico seleccionado del grupo que comprende: KLK2, PMEPA1, TMPRSS2, NKX3-1, ABCC4, KLK3, FKBP5, ELL2, UGT2B15, DHCR24, PPAP2A, NDRG1, LRIG1, CREB3L4, LCP1, GUCY1A3, AR y EAF2.

8. El método de la reivindicación 4, en el que la inferencia se basa adicionalmente en los niveles (20) de expresión de al menos un gen objetivo de la ruta Wnt medidos en la muestra extraída del tejido del sujeto médico seleccionados del grupo que comprende: NKD1, OAT, FAT1, LEF1, GLUL, REG1B, TCF7L2, COL18A1, BMP7, SLC1A2, ADRA2C, PPARG, DKK1, HNF1A y LECT2.

9. El método de la reivindicación 5, en el que la inferencia se basa adicionalmente en los niveles (20) de expresión de al menos un gen objetivo de la ruta ER medidos en la muestra extraída del tejido del sujeto médico seleccionados del grupo que comprende: AP1B1, ATP5J, COL18A1, COX7A2L, EBAG9, ESR1, HSPB1, IGFBP4, KRT19, MYC, NDUFV3, PISD, PRDM15, PTMA, RARA, SOD1 y TRIM25.

10. El método de la reivindicación 6, en el que la inferencia se basa adicionalmente en los niveles (20) de expresión de al menos un gen objetivo de la ruta Hedgehog medidos en la muestra extraída del tejido del sujeto médico seleccionados del grupo que comprende: BCL2, FOXA2, FOXF1, H19, HHIP, IL1R2, JAG2, JUP, MIF, MYLK, NKX2-2, NKX2-8, PITRM1 y TOM1.

11. El método de la reivindicación 7, en el que la inferencia se basa adicionalmente en los niveles (20) de expresión de al menos un gen objetivo de la ruta AR medidos en la muestra extraída del tejido del sujeto médico seleccionados del grupo que comprende: APP, NTS, PLAU, CDKN1A, DRG1, FGF8, IGF1, PRKACB, PTPN1, SGK1 y TACC2.

12. El método de una cualquiera de las reivindicaciones 1-11 que comprende:

inferir la actividad de una ruta Wnt en el tejido de un sujeto médico con base en al menos en los niveles (20) de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta Wnt medidos en una muestra extraída del tejido del sujeto médico

y/o

inferir la actividad de una ruta ER en el tejido de un sujeto médico con base en al menos en los niveles (20) de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta ER medidos en una muestra extraída del tejido del sujeto médico

y/o

inferir la actividad de una ruta Hedgehog en el tejido de un sujeto médico con base en al menos en los niveles (20) de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta Hedgehog medidos en una muestra extraída del tejido del sujeto médico,

y/o

inferir la actividad de una ruta AR en el tejido de un sujeto médico con base en al menos en los niveles (20) de expresión de dos, tres o más genes objetivo de un conjunto de genes objetivo de la ruta AR medidos en una muestra extraída del tejido del sujeto médico.

13. El método de la reivindicación 12, en el que

el conjunto de genes objetivo de la ruta Wnt incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende: KIAA1199, AXIN2, RNF43, TBX3, TDGF1, SOX9, ASCL2, IL8, SP5, ZNRF3, KLF6, CCND1, DEFA6 y FZD7,

y/o

el conjunto de genes objetivo de la ruta ER incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende: CDH26, SGK3, PGR, GREB1, CA12, XBP1, CELSR2, WISP2, DSCAM, ERBB2, CTSD, TFF1 y NRIP1,

y/o

el conjunto de genes objetivo de la ruta Hedgehog incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende: GLI1, PTCH1, PTCH2, IGFBP6, SPP1, CCND2, FST, FOXL1, CFLAR, TSC22D1, RAB34, S100A9, S100A7, MYCN, FOXM1, GLI3, TCEA2, FYN y CTSL1,

5 y/o

el conjunto de genes objetivo de la ruta AR incluye al menos nueve, preferiblemente todos los genes objetivo seleccionados del grupo que comprende: KLK2, PMEPA1, TMPRSS2, NKX3-1, ABCC4, KLK3, FKBP5, ELL2, UGT2B15, DHCR24, PPAP2A, NDRG1, LRIG1, CREB3L4, LCP1, GUCY1A3, AR y EAF2.

10

14. El método de la reivindicación 13, en el que

el conjunto de genes objetivo de la ruta Wnt incluye además al menos un gen objetivo seleccionado del grupo que comprende: NKD1, OAT, FAT1, LEF1, GLUL, REG1B, TCF7L2, COL18A1, BMP7, SLC1A2, ADRA2C, PPARG, DKK1, HNF1A, y LECT2,

15

y/o

el conjunto de genes objetivo de la ruta ER incluye además al menos un gen objetivo seleccionado del grupo que comprende: AP1B1, ATP5J, COL18A1, COX7A2L, EBAG9, ESR1, HSPB1, IGFBP4, KRT19, MYC, NDUFV3, PISD, PRDM15, PTMA, RARA, SOD1 y TRIM25,

20

y/o

el conjunto de genes objetivo de la ruta Hedgehog incluye además al menos un gen objetivo seleccionado del grupo que comprende: BCL2, FOXA2, FOXF1, H19, HHIP, IL1R2, JAG2, JUP, MIF, MYLK, NKX2-2, NKX2-8, PITRM1 y TOM1,

25

y/o

el conjunto de genes objetivo de la ruta AR incluye además al menos un gen objetivo seleccionado del grupo que comprende: APP, NTS, PLAU, CDKN1A, DRG1, FGF8, IGF1, PRKACB, PTPN1, SGK1 y TACC2.

30

15. El método de cualquiera de las reivindicaciones 1-14, que comprende además:

35

determinar si las rutas de señalización celular está funcionando anormalmente en el tejido del sujeto médico con base en la actividad inferida de las rutas de señalización celular en el tejido del sujeto médico.

16. Un aparato que comprende un procesador (12) digital configurado para llevar a cabo un método según se expone en una cualquiera de las reivindicaciones 1-15 o un medio de almacenamiento no transitorio que almacena instrucciones que son ejecutables por un dispositivo (12) de procesamiento digital para realizar un método como el expuesto en una cualquiera de las reivindicaciones 1-15 o un programa de ordenador que comprende medios de código de programa para hacer que un dispositivo (12) de procesamiento digital realice un método como el expuesto en una cualquiera de las reivindicaciones 1-15.

40

45

FIG. 1

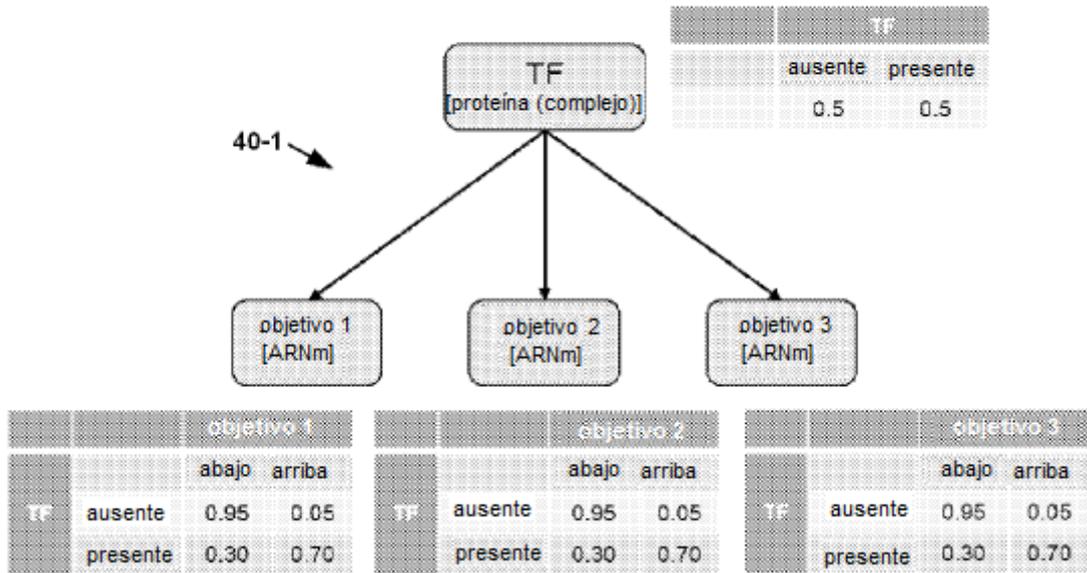


FIG. 2

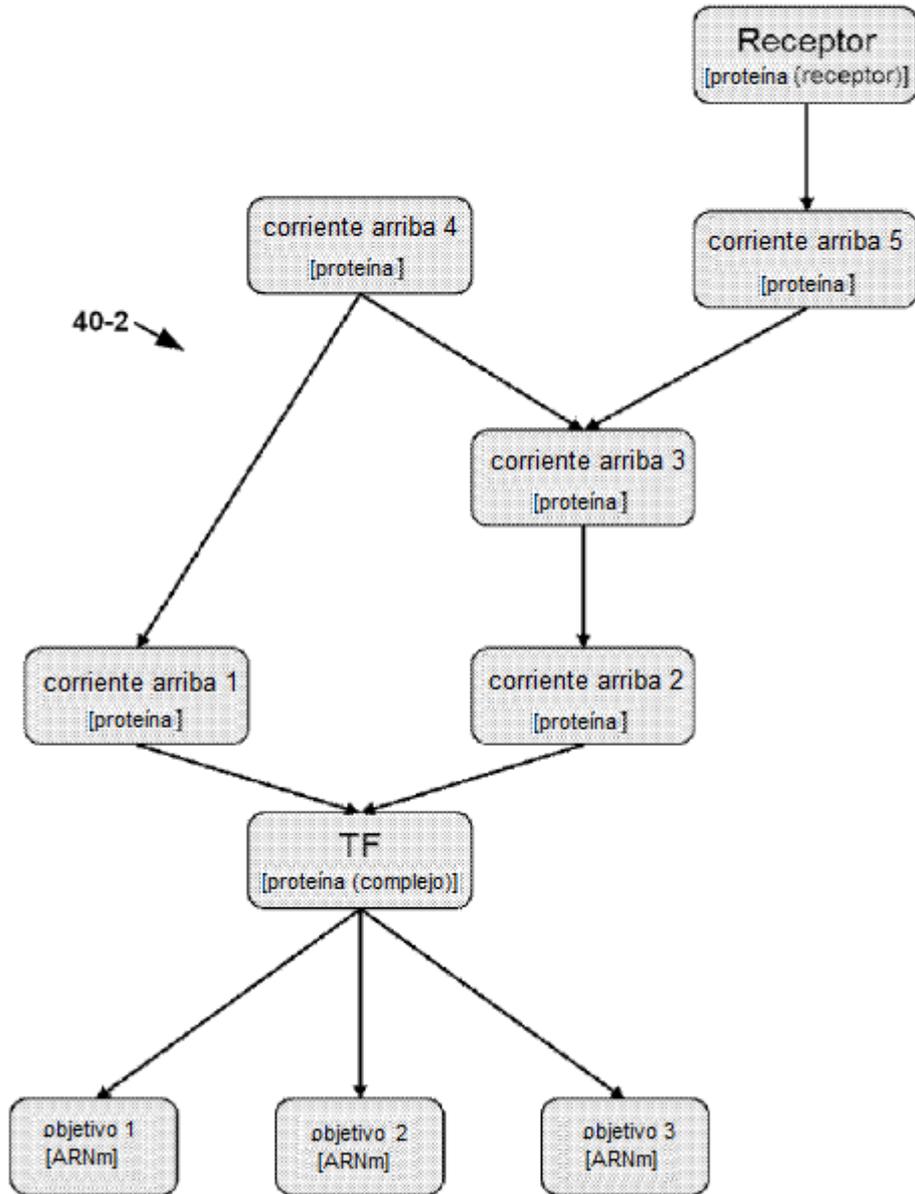
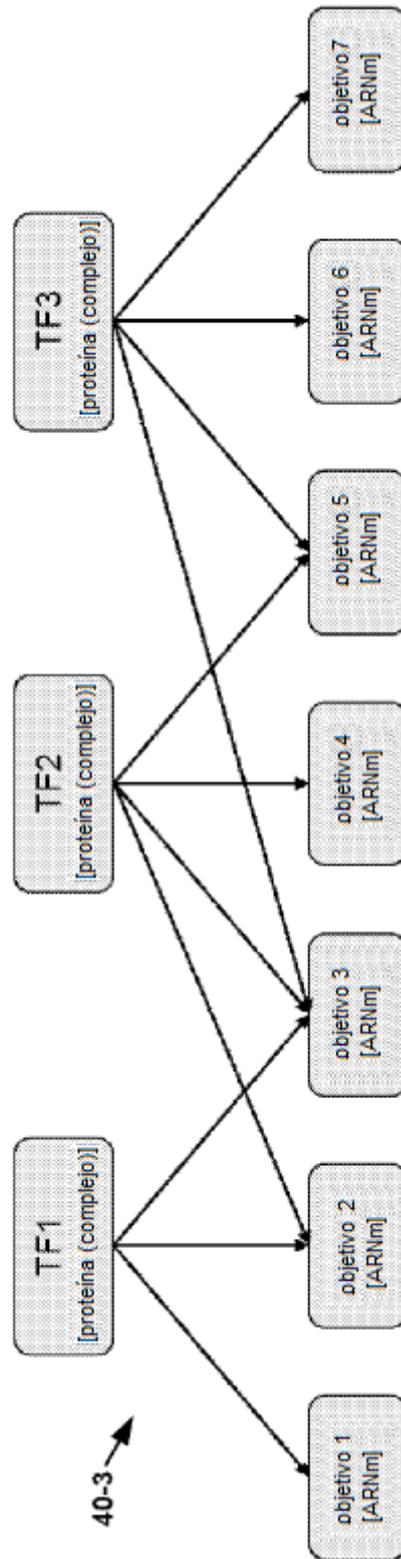


FIG. 3



40-3 →

FIG. 4

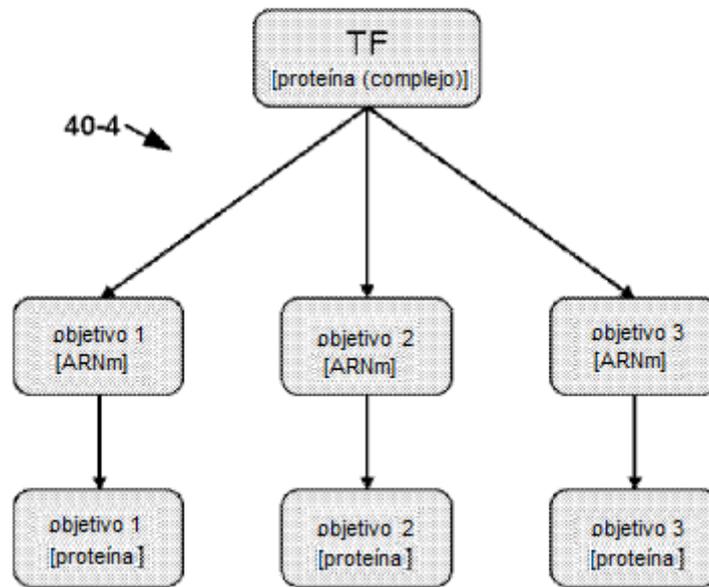


FIG. 5

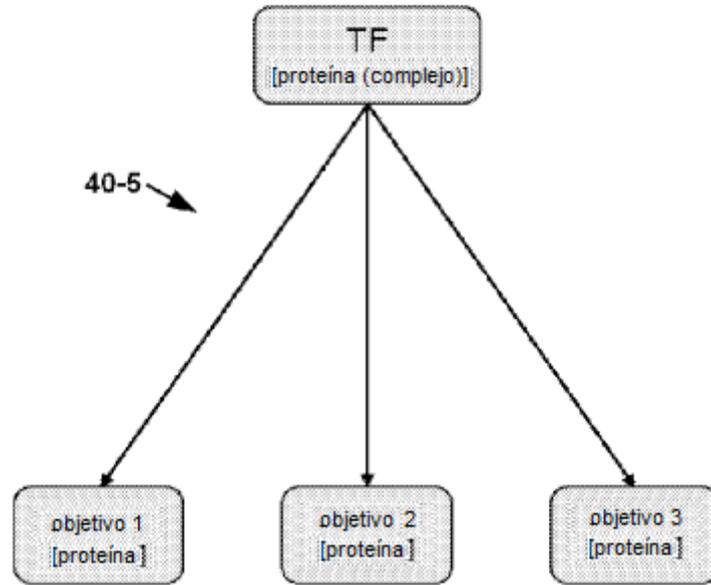


FIG. 6

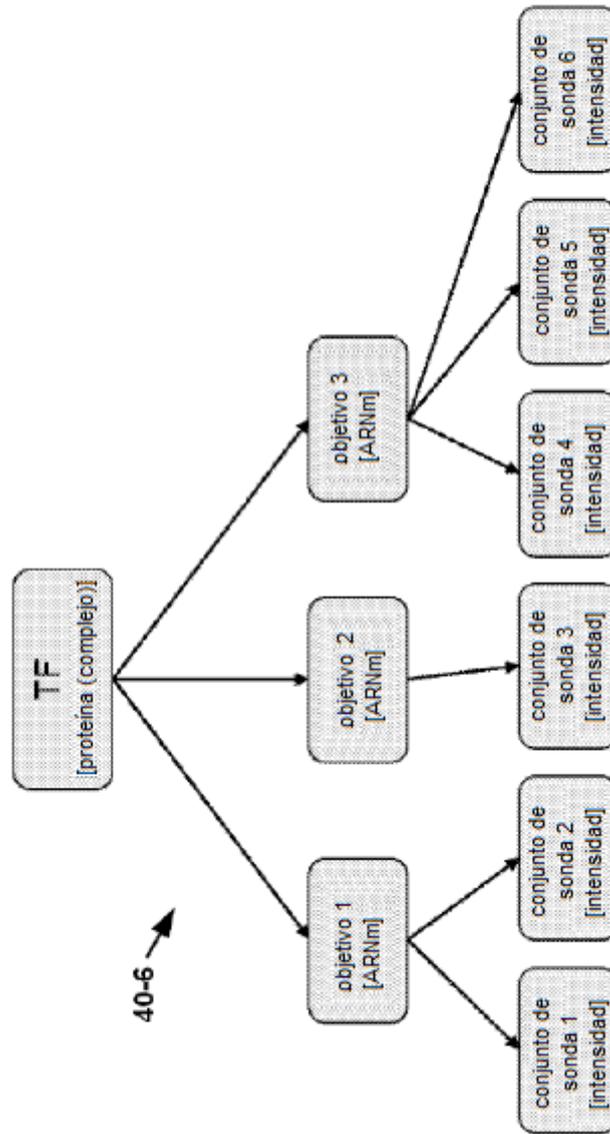


FIG. 7

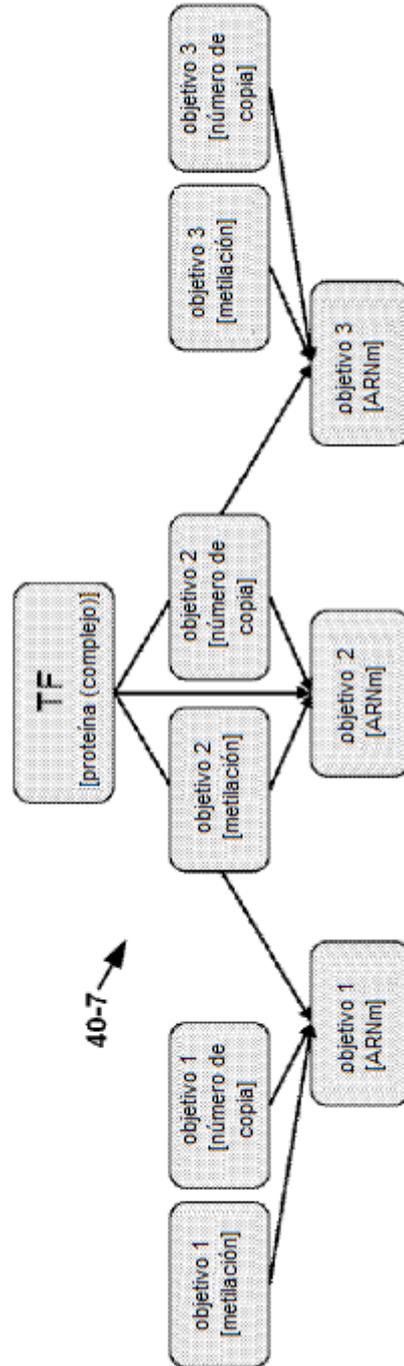




FIG. 9

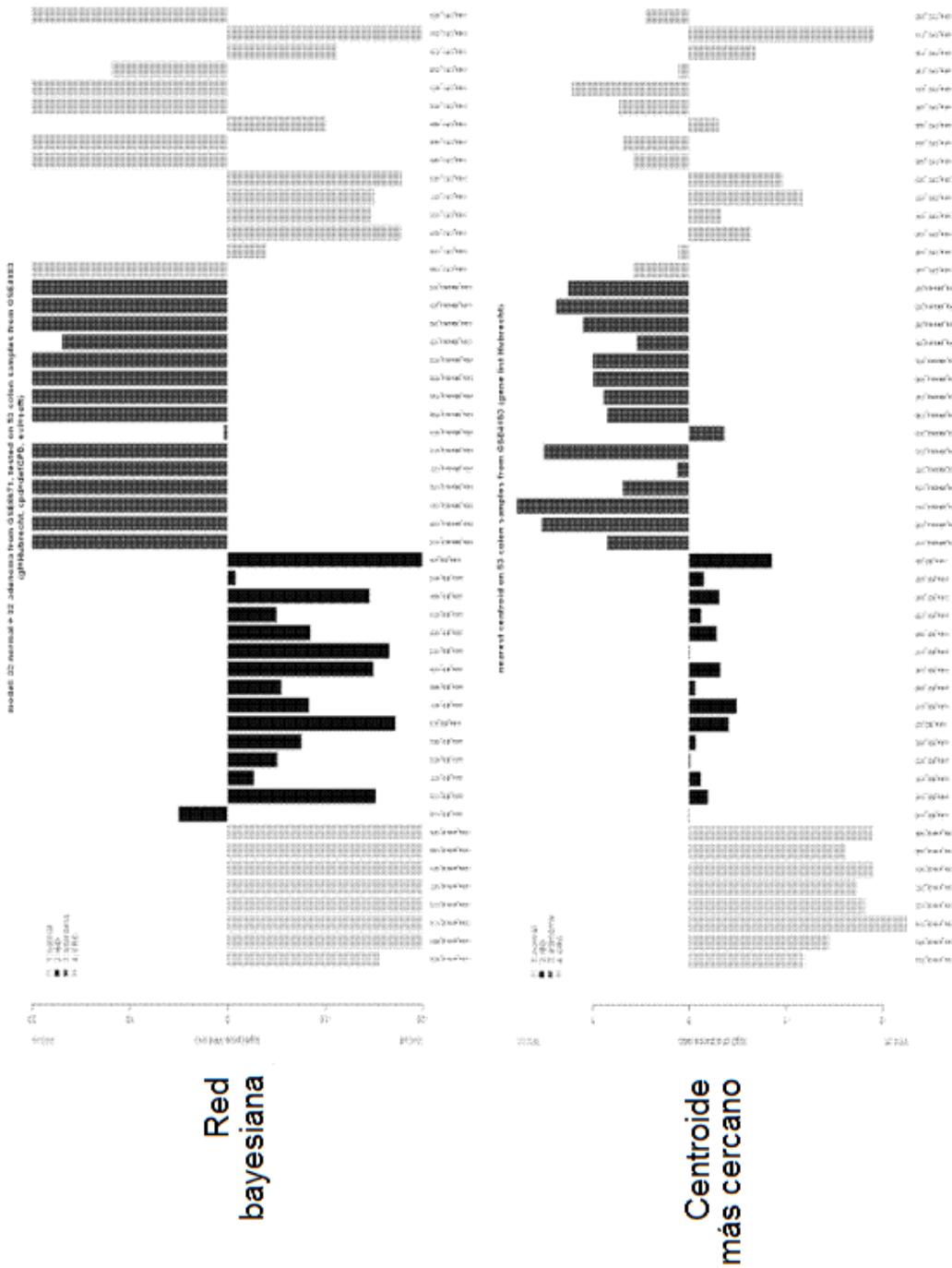


FIG. 10

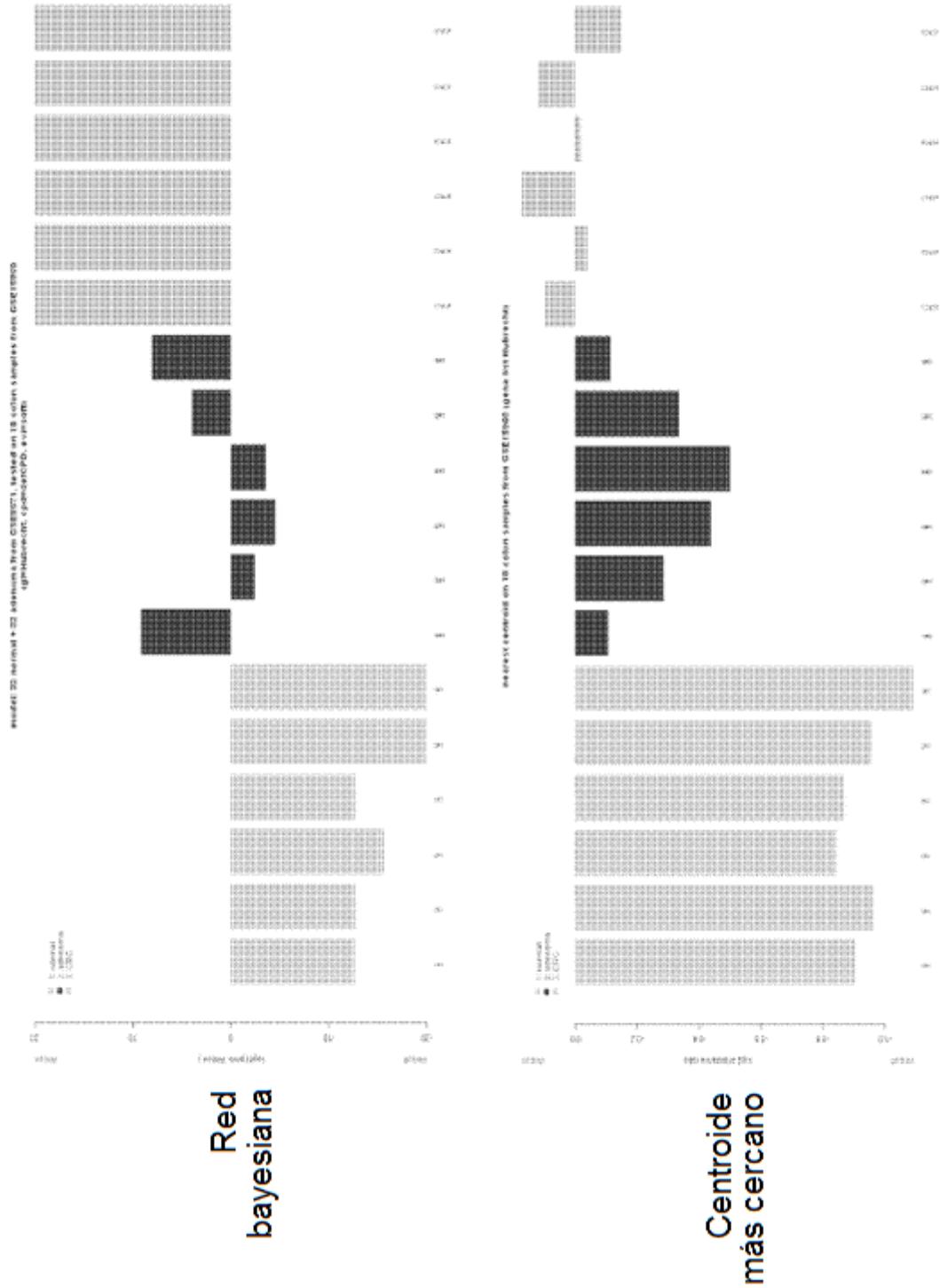
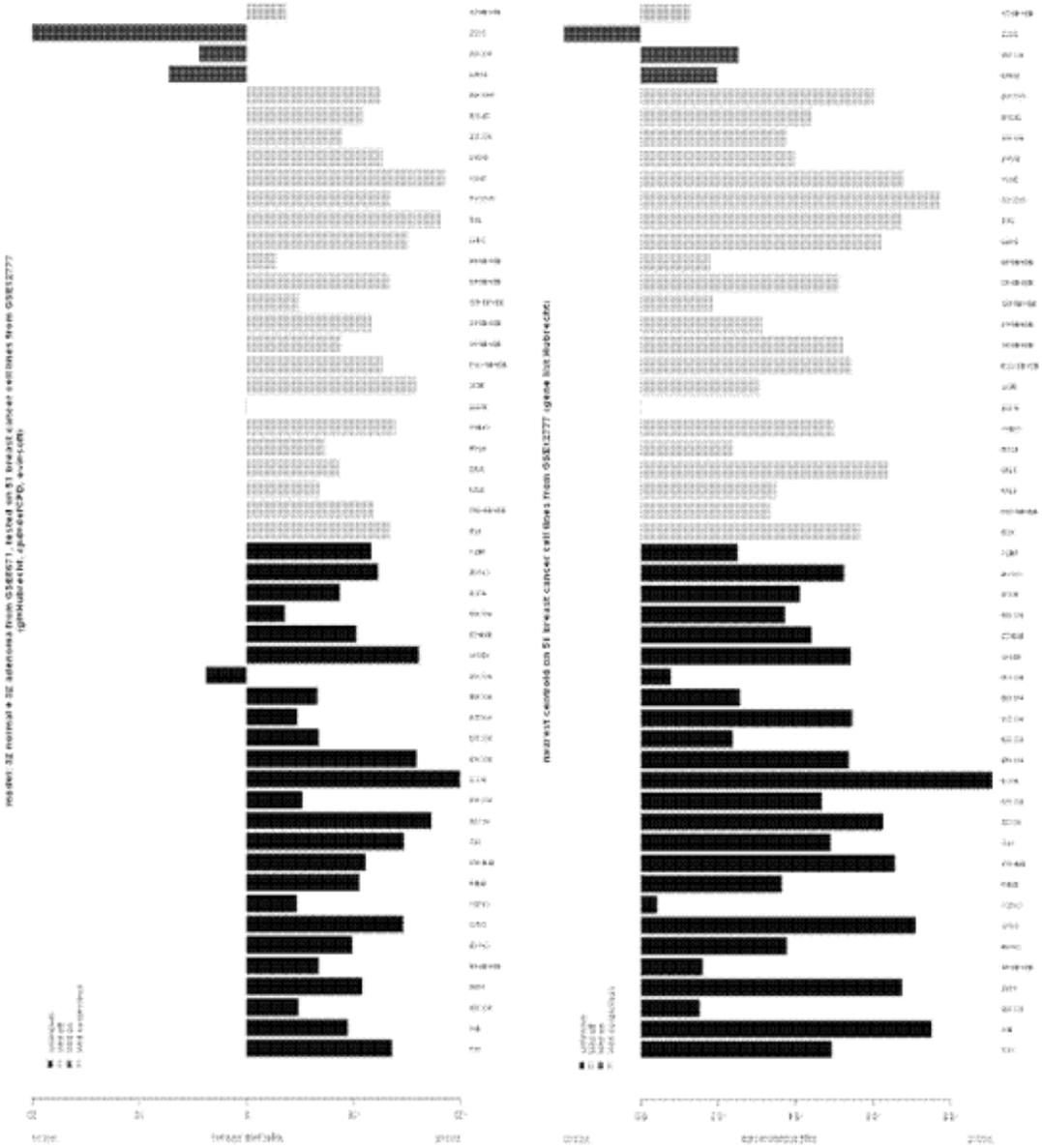


FIG. 11



Red bayesiana

Centroide más cercano

FIG. 12

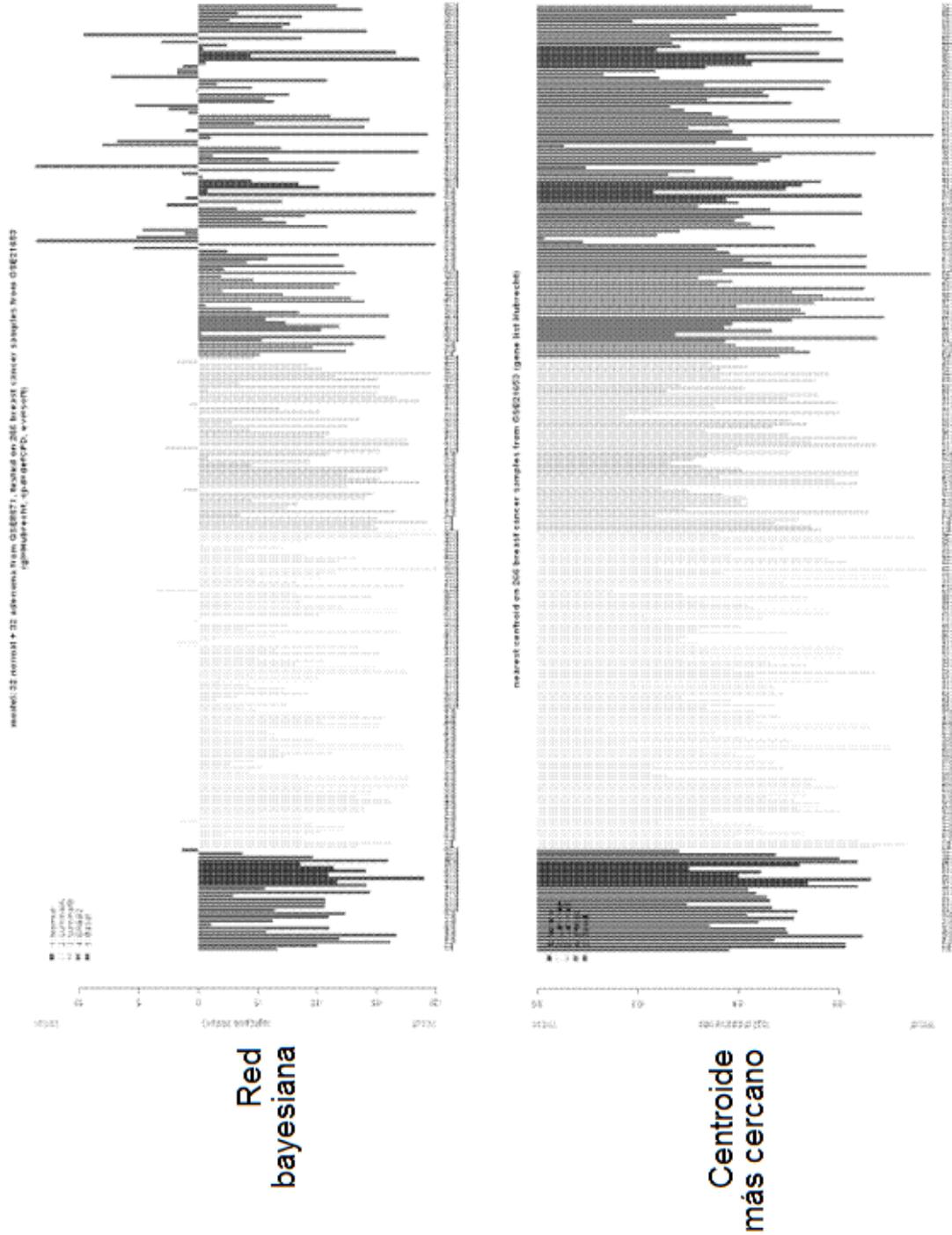


FIG. 13

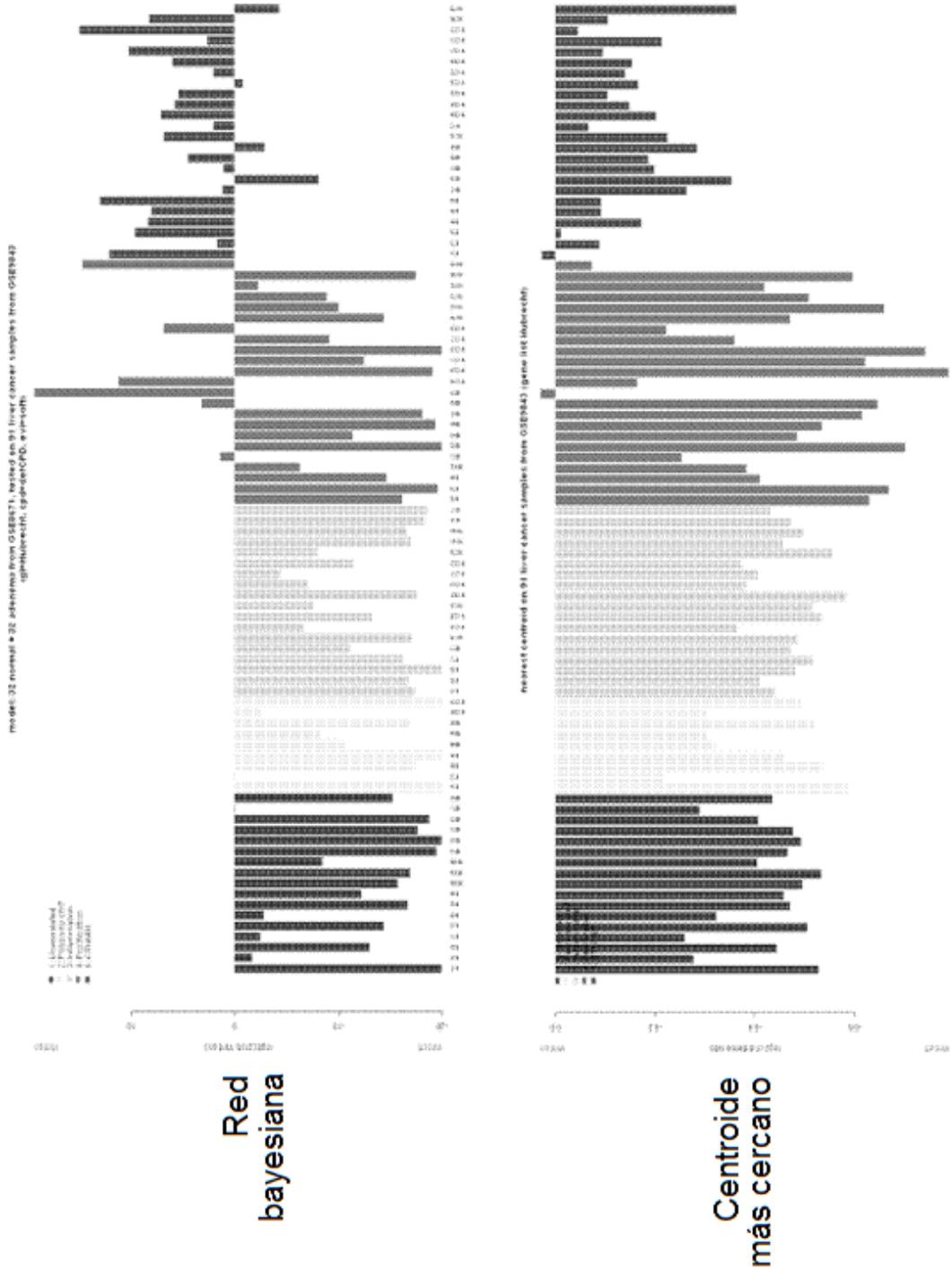


FIG. 14  
Lista de evidencias científicas



FIG. 14 (CONTINUADO)  
**Lista amplia de literatura**

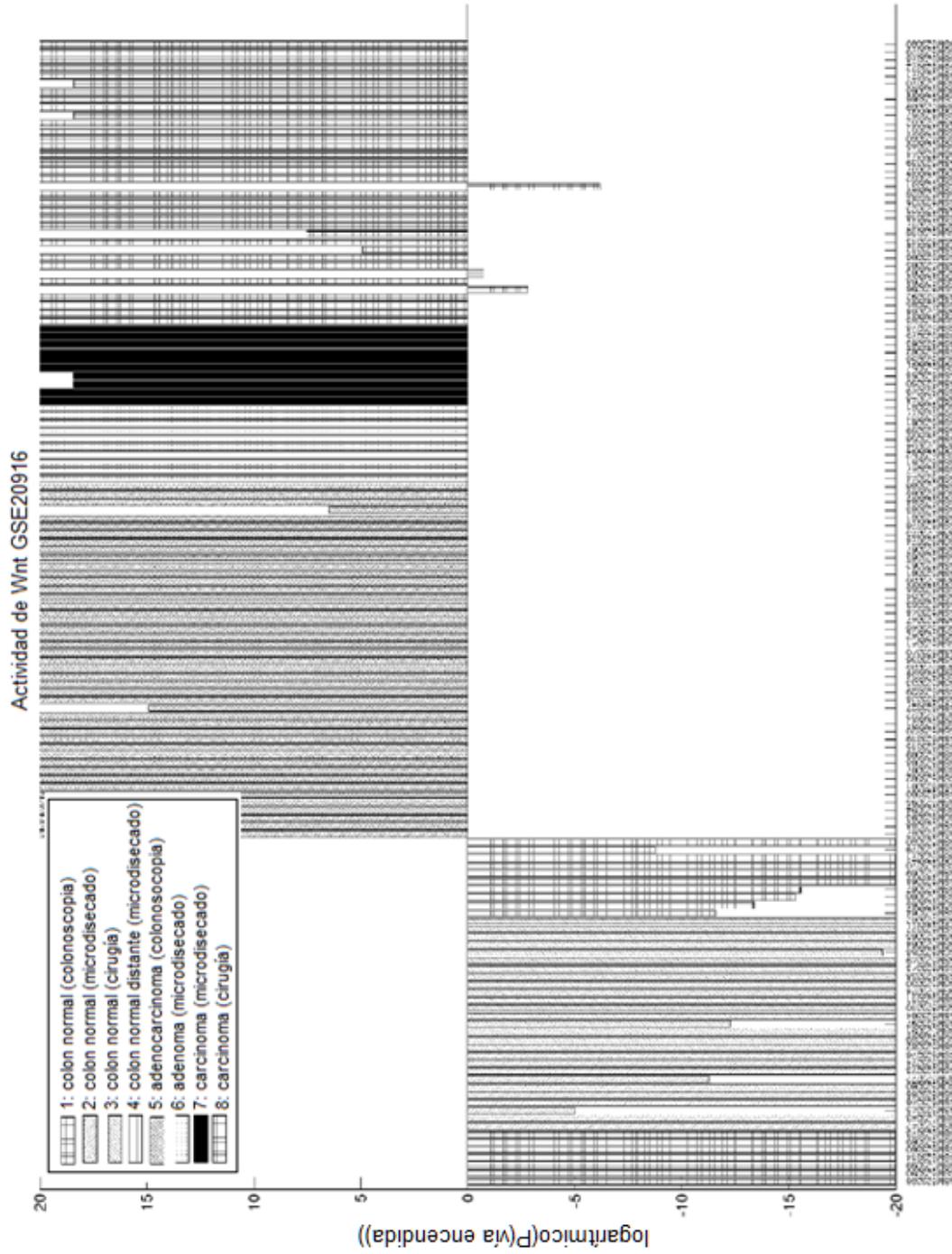


FIG. 15  
**Lista de evidencias científicas**

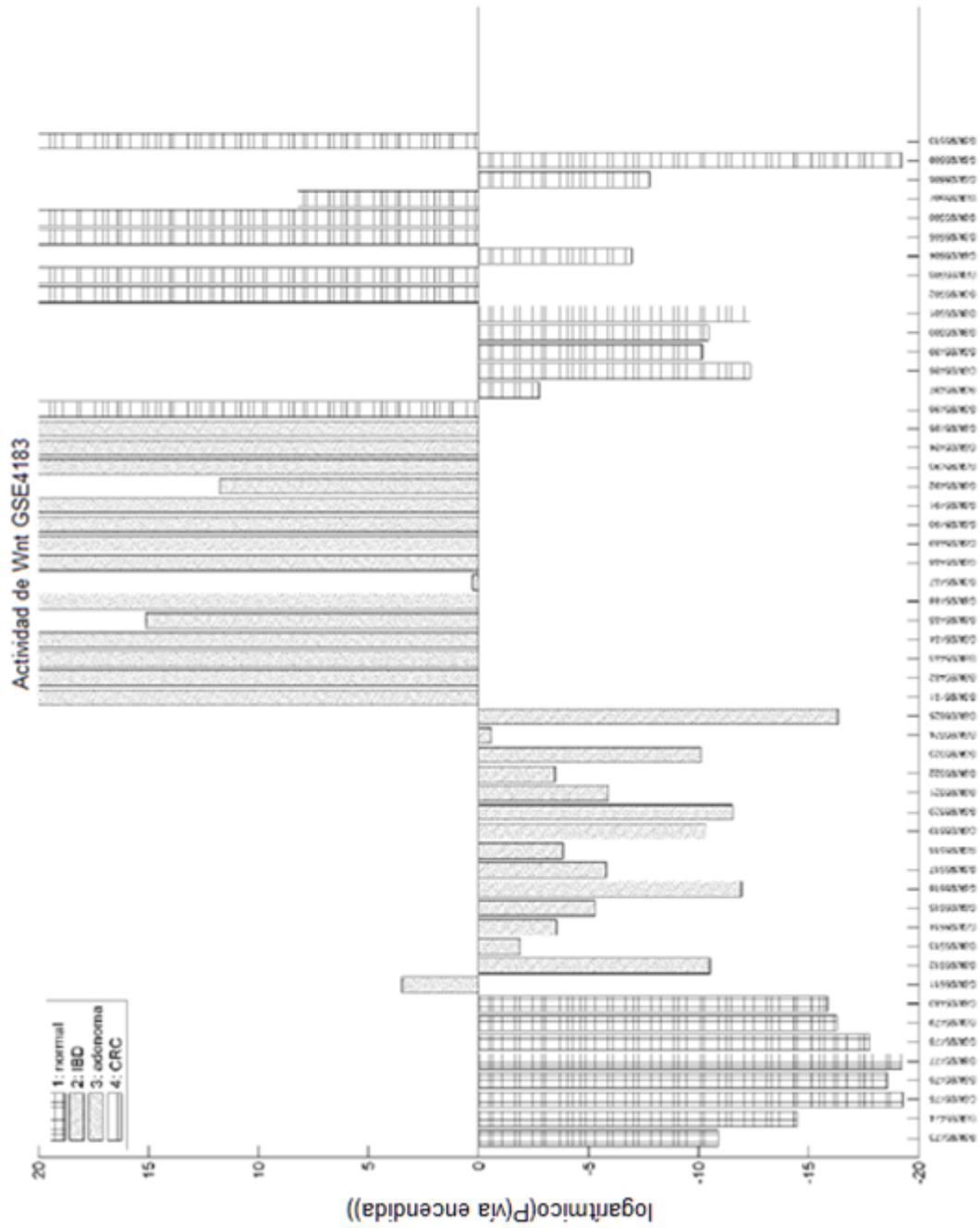


FIG. 15 (CONTINUADO)  
 Lista amplia de literatura

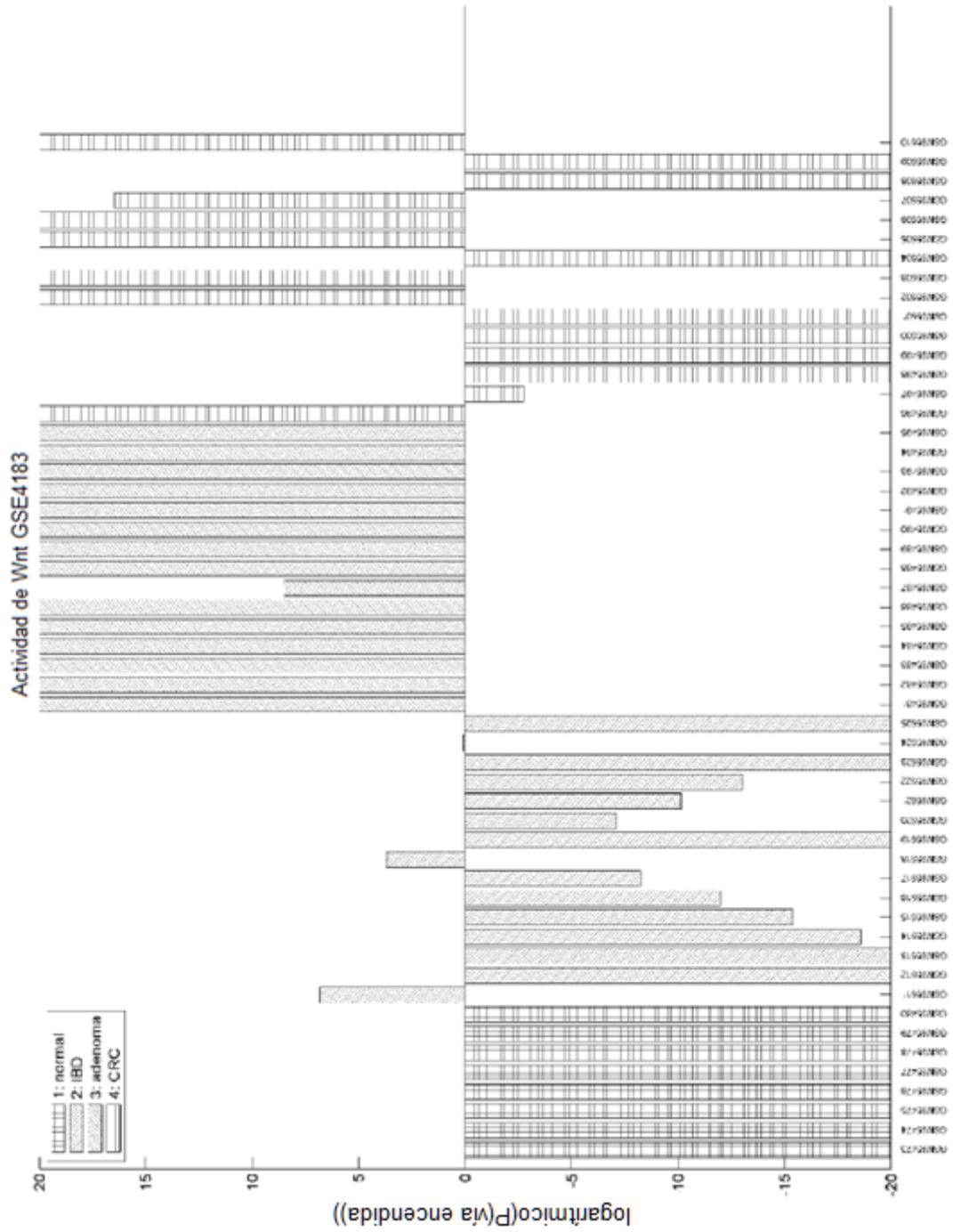


FIG. 16

**Lista de evidencias científicas**

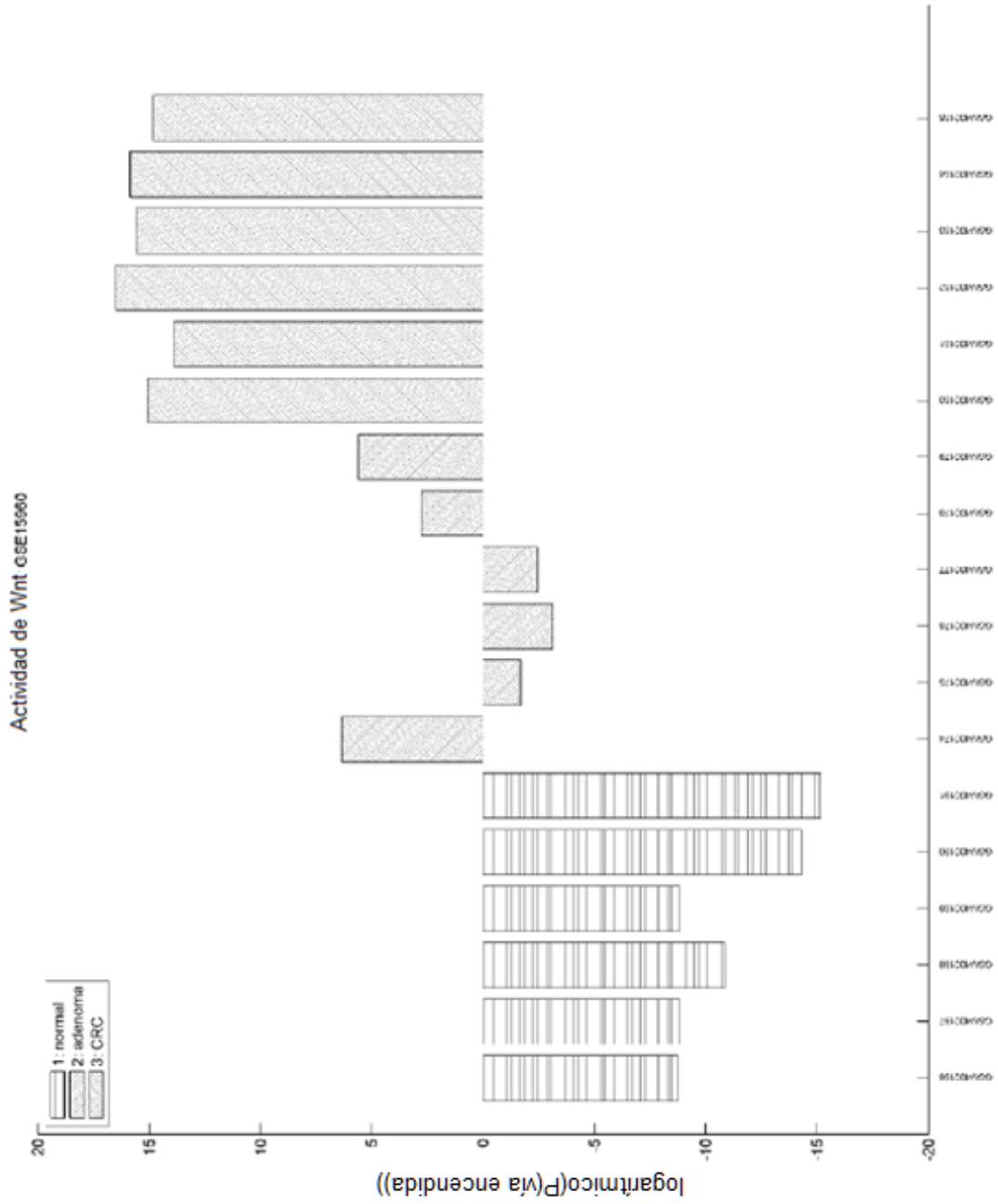


FIG. 16 (CONTINUADO)  
**Lista amplia de literatura**

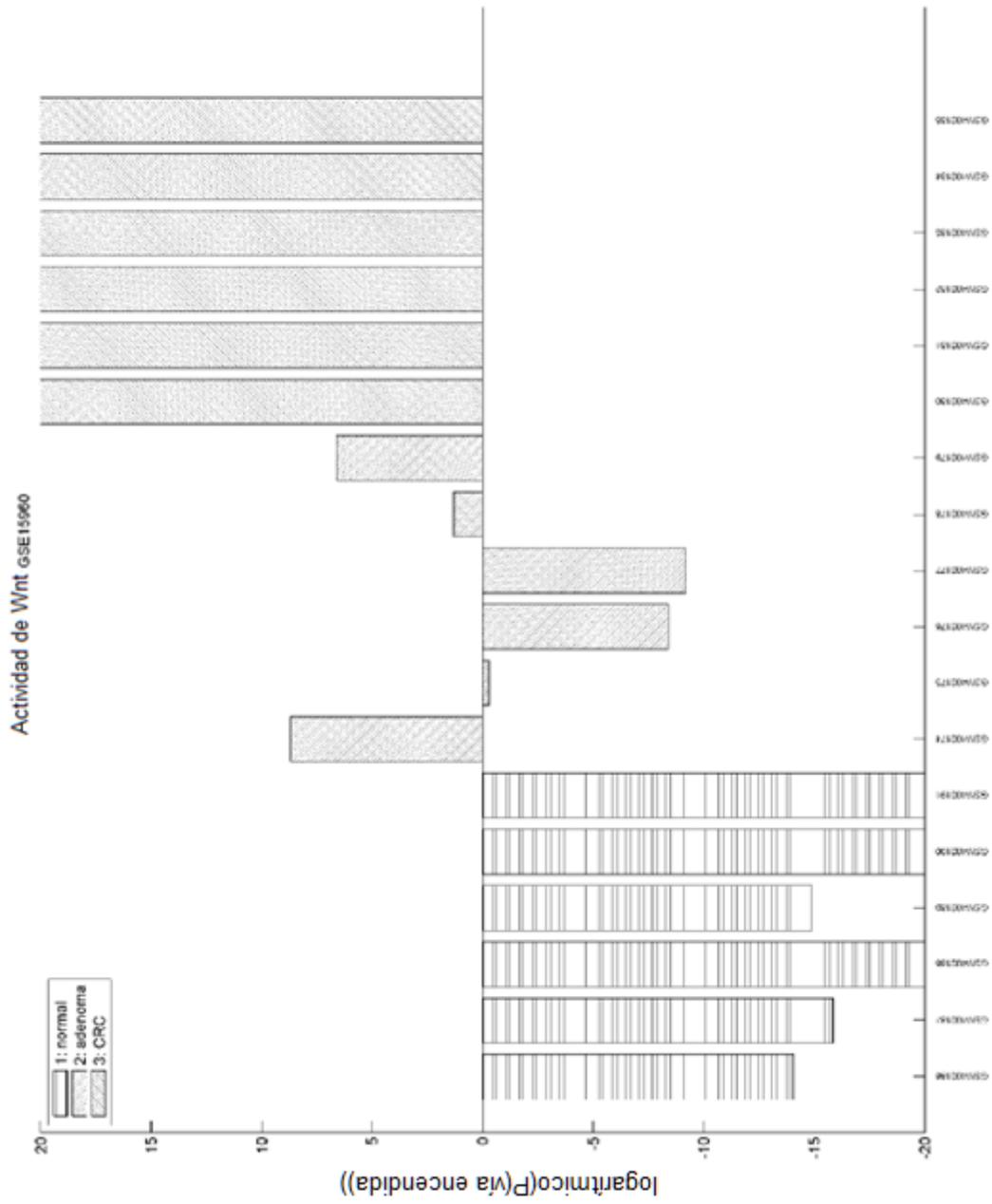












FIG. 19 (CONTINUADO)  
**Lista amplia de literatura**

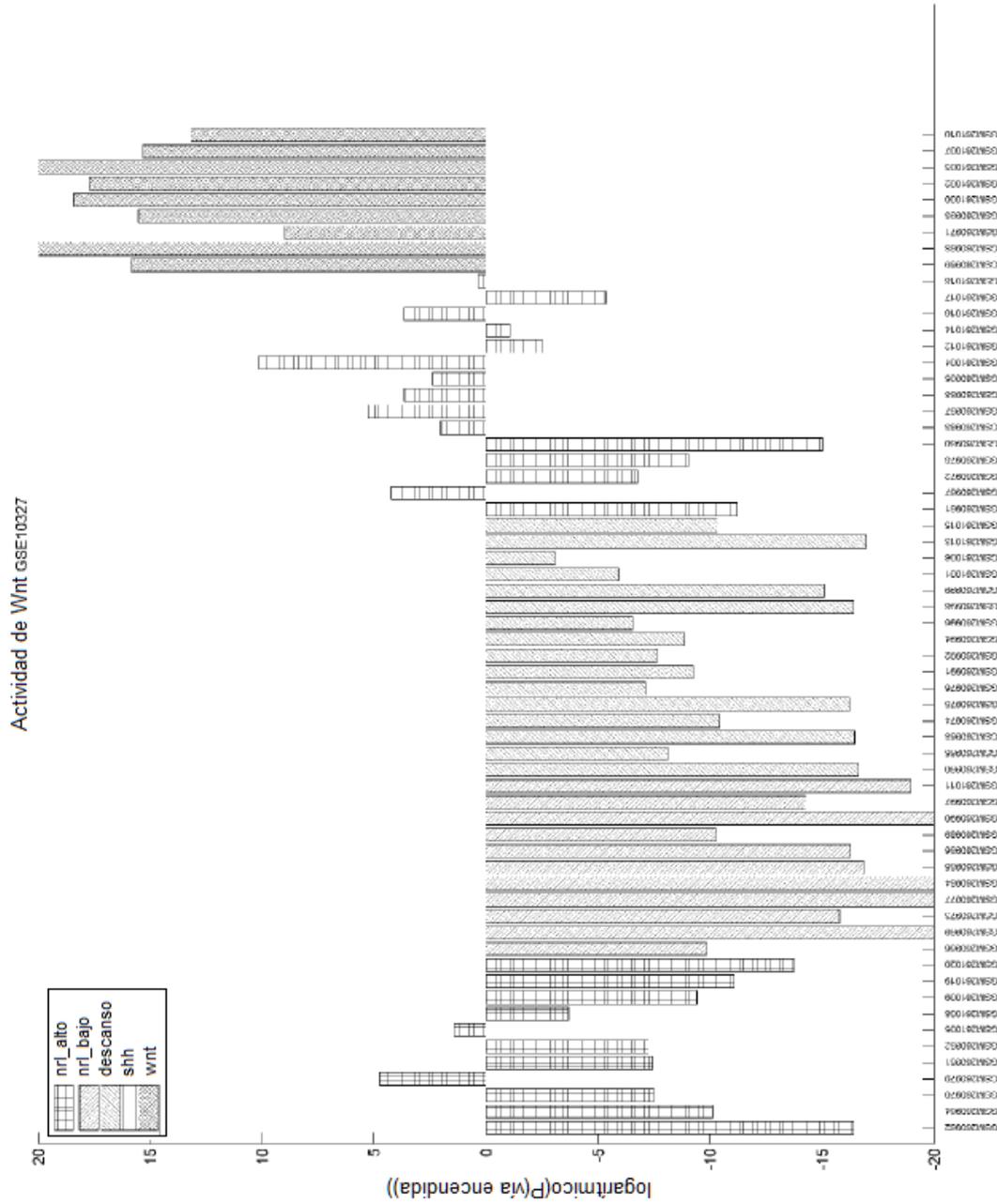


FIG. 20

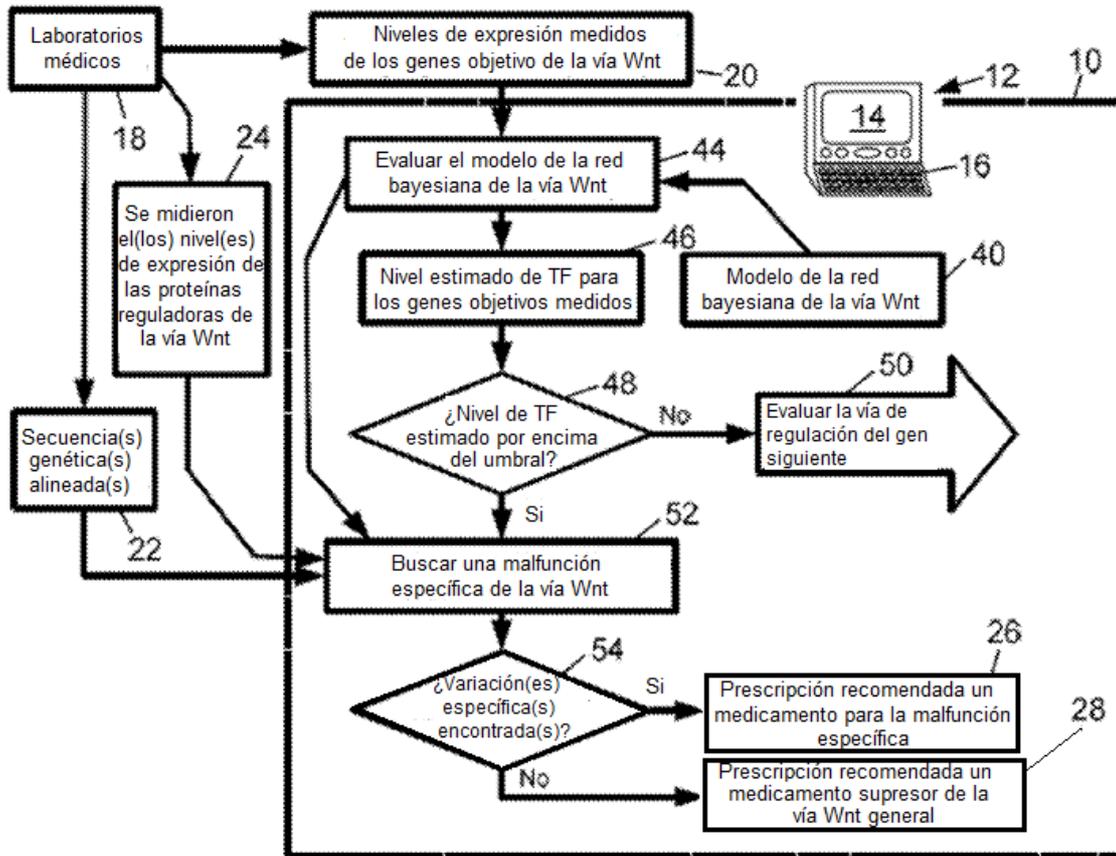






FIG. 23

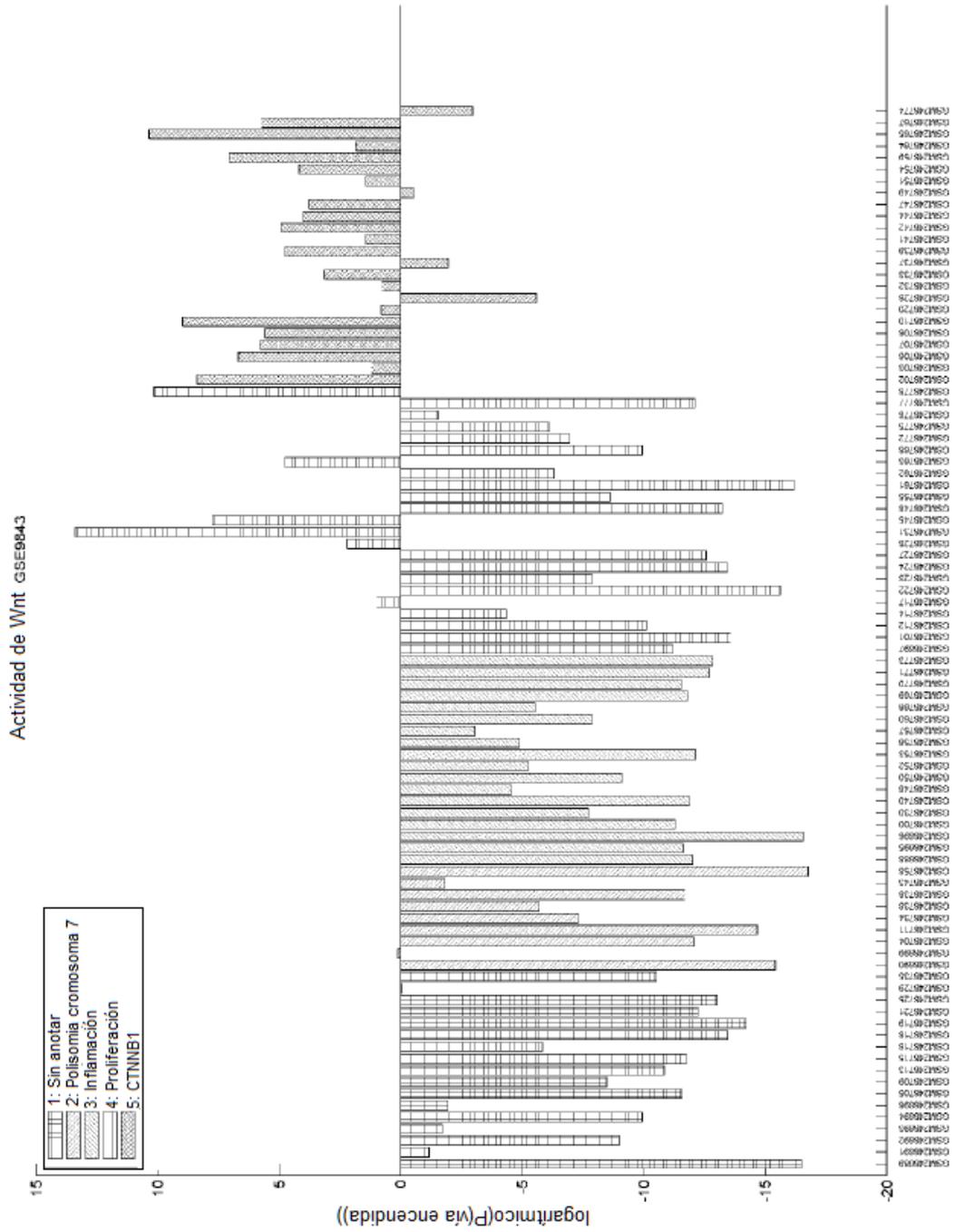










FIG. 28

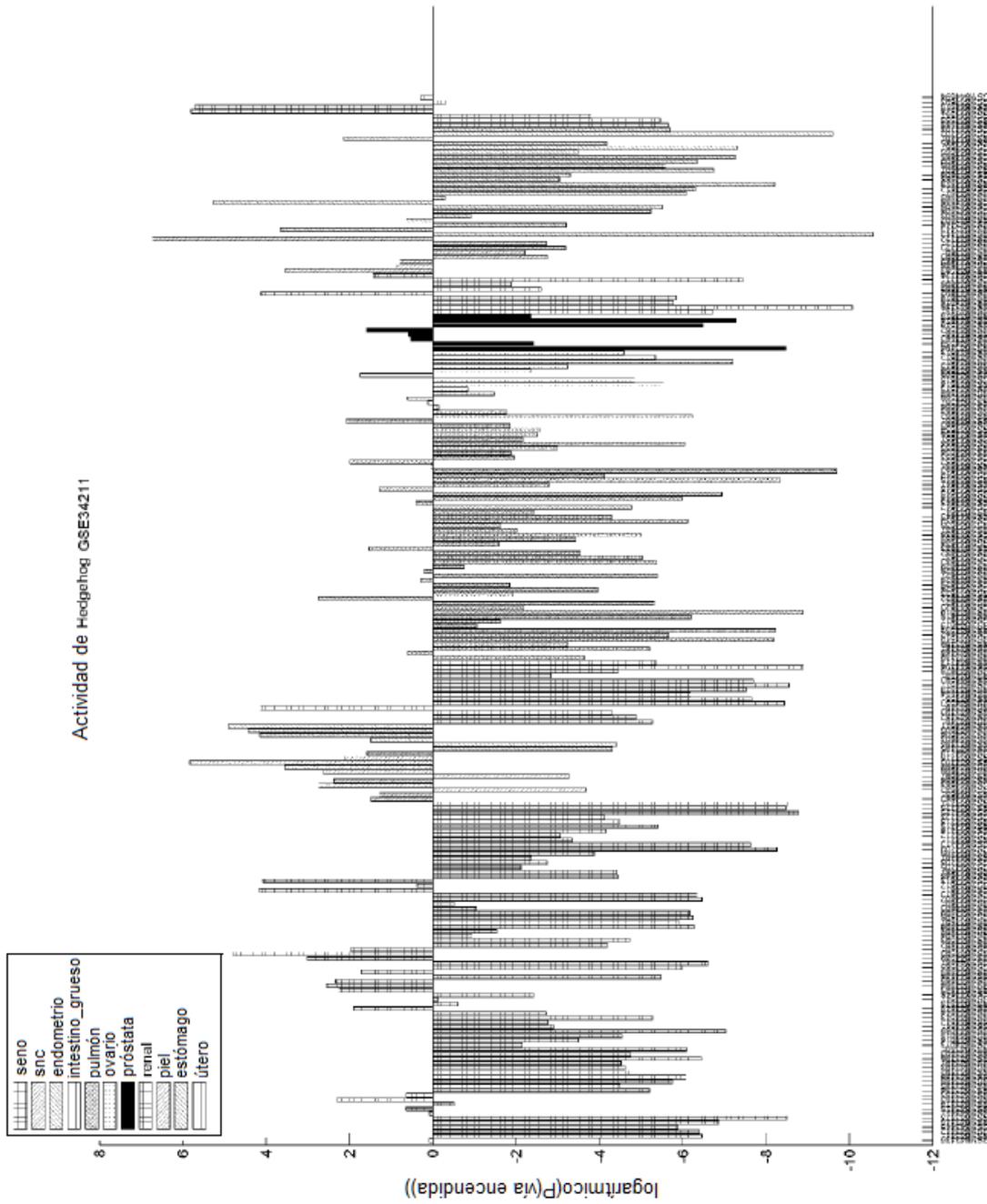




FIG. 30

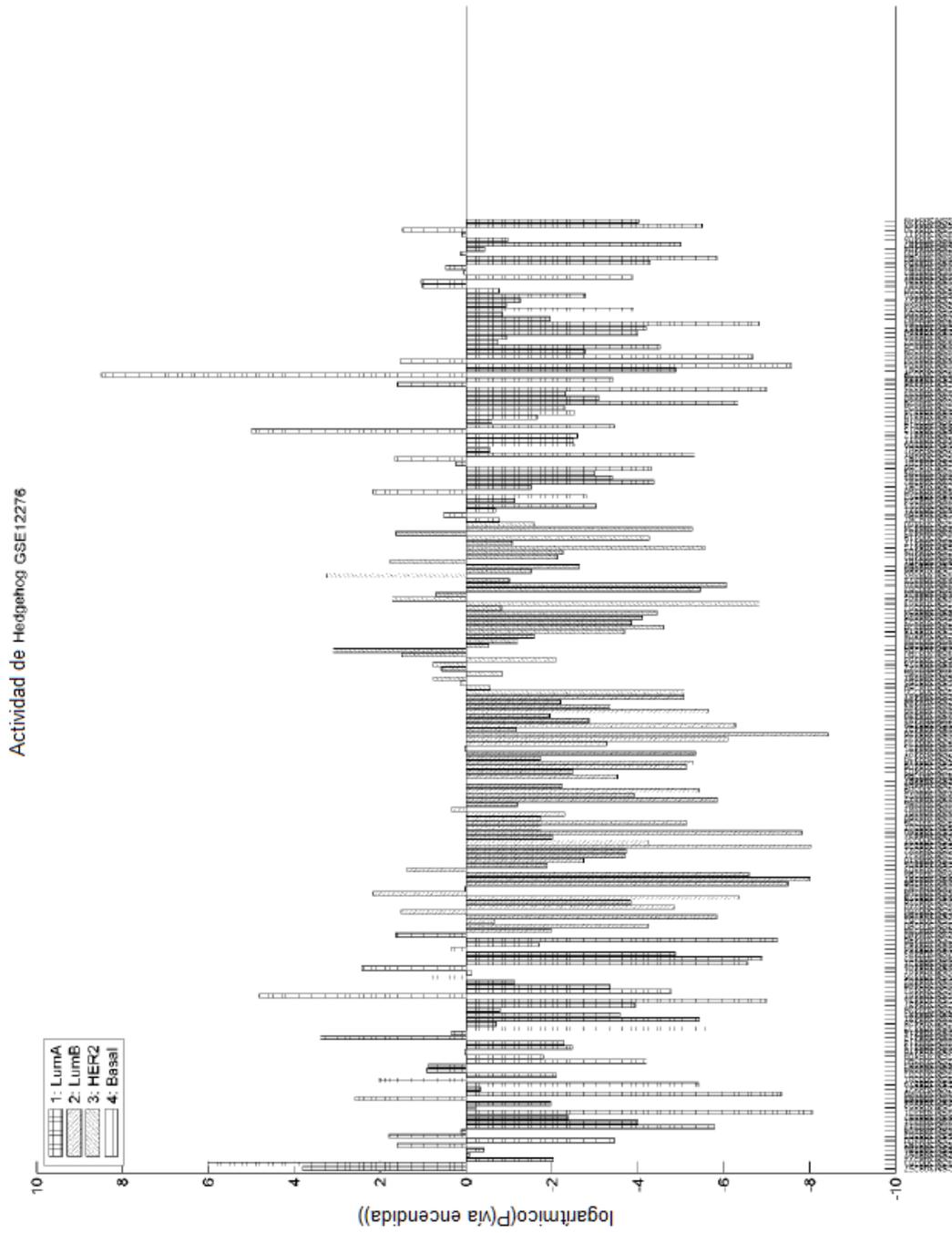


FIG. 31

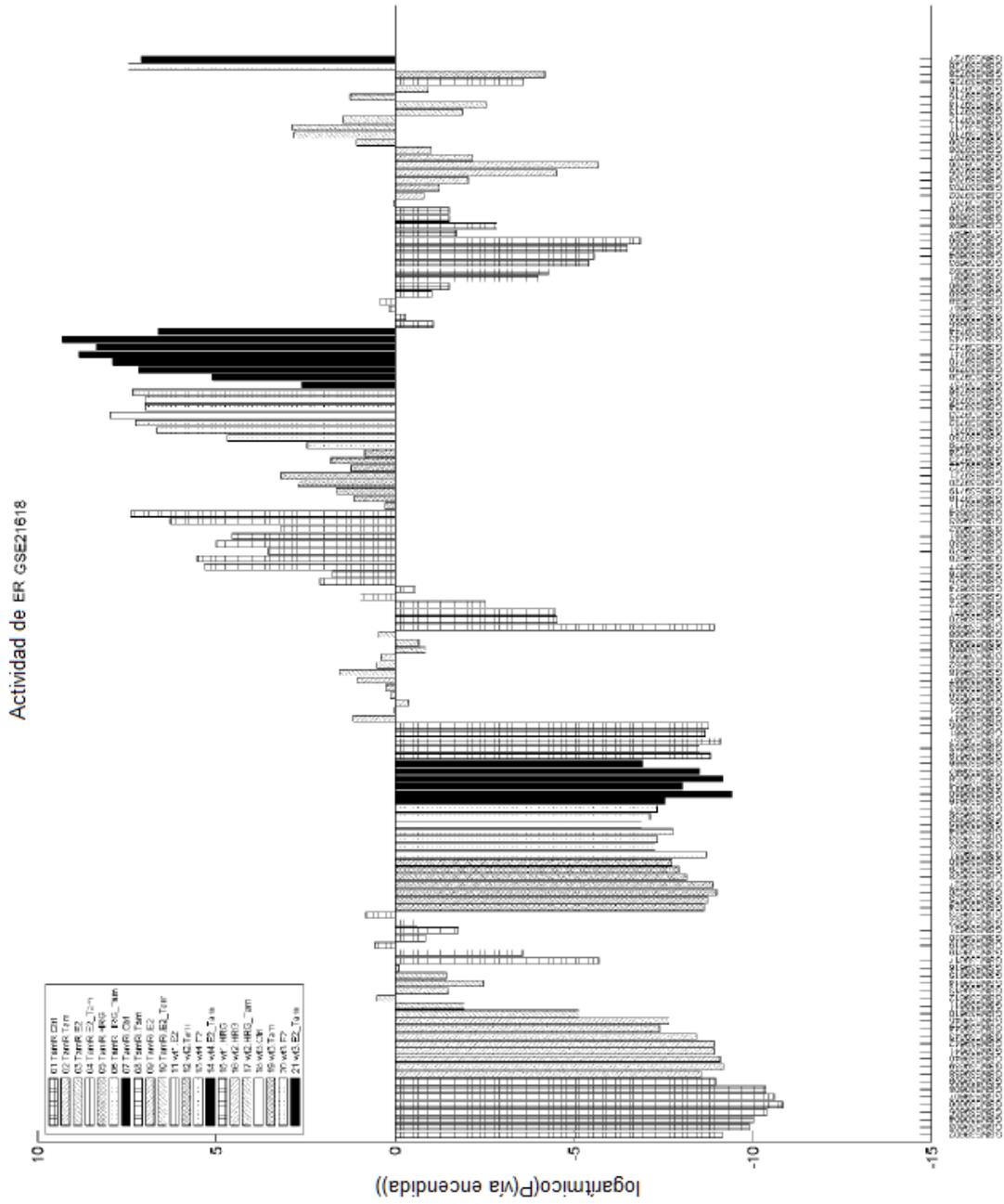


FIG. 32

Modelo: 4 estrógenos privados y 4 estrógenos estimulados con muestras MCF7 de GSE6587, probado en 12 muestras MCF7 con serie temporal de estimulación de estrógenos de GSE11324 (g)ER lista (p=defCPD, e(v)= suave)

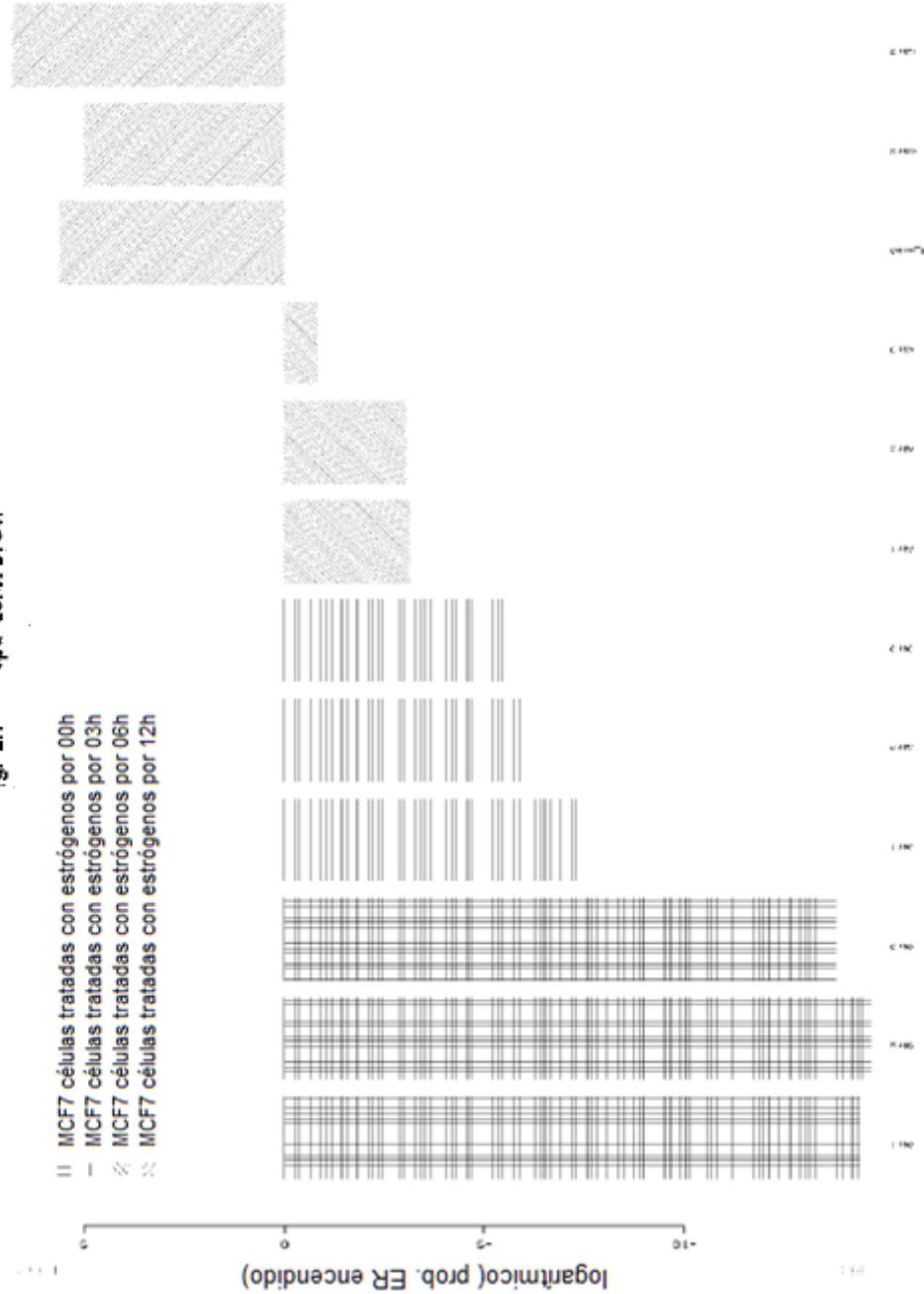


FIG. 33

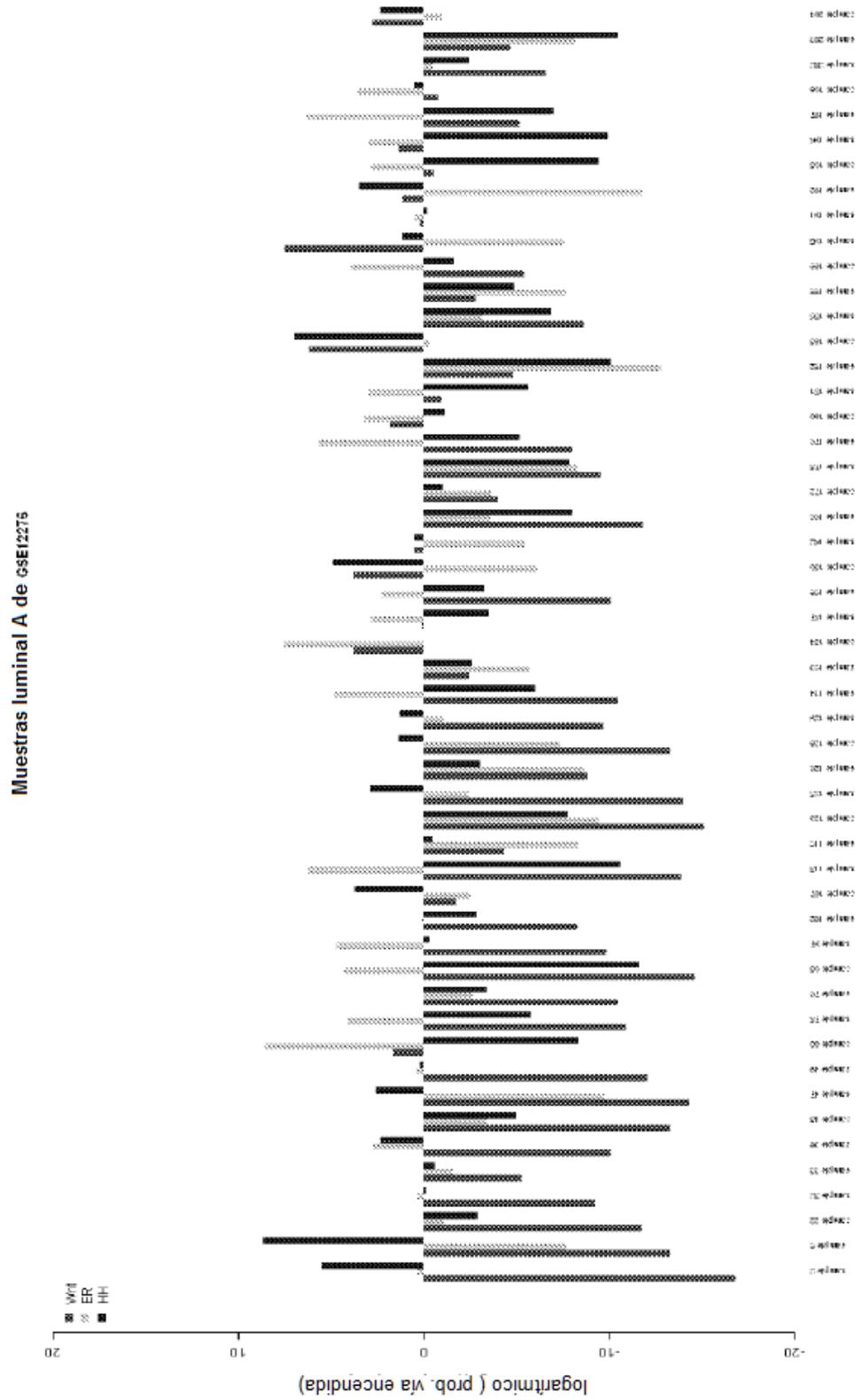


FIG. 34

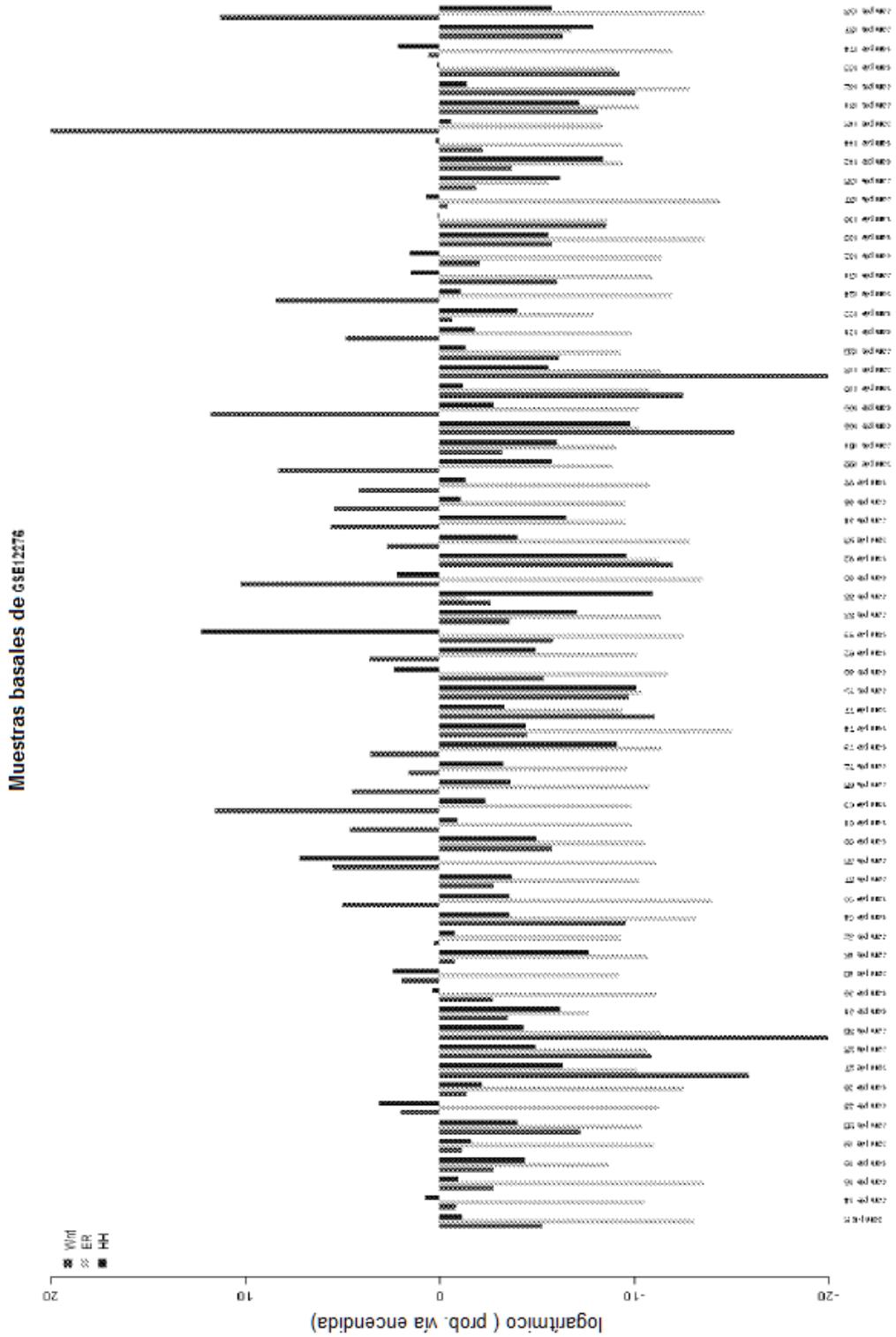


FIG. 35

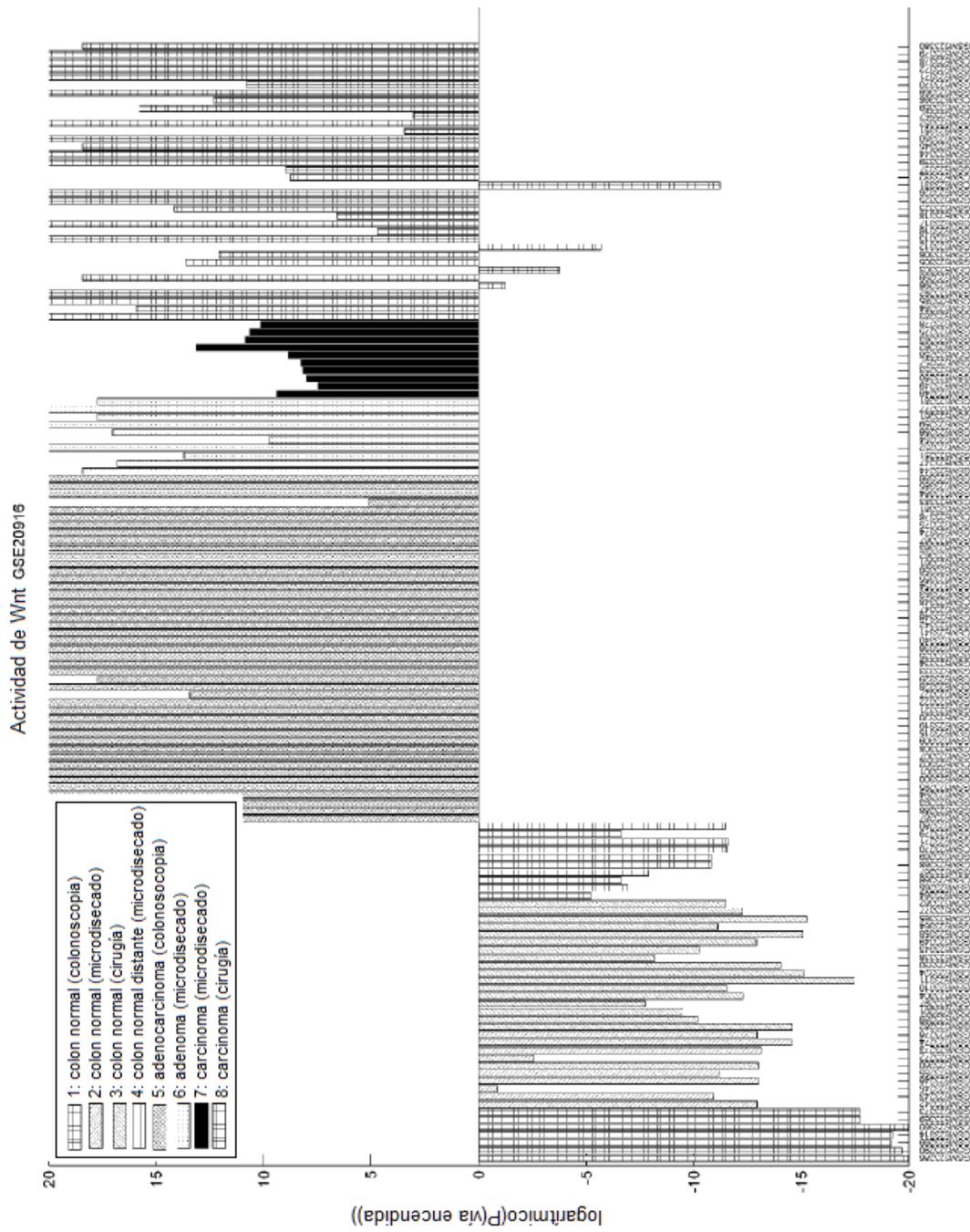




FIG. 37

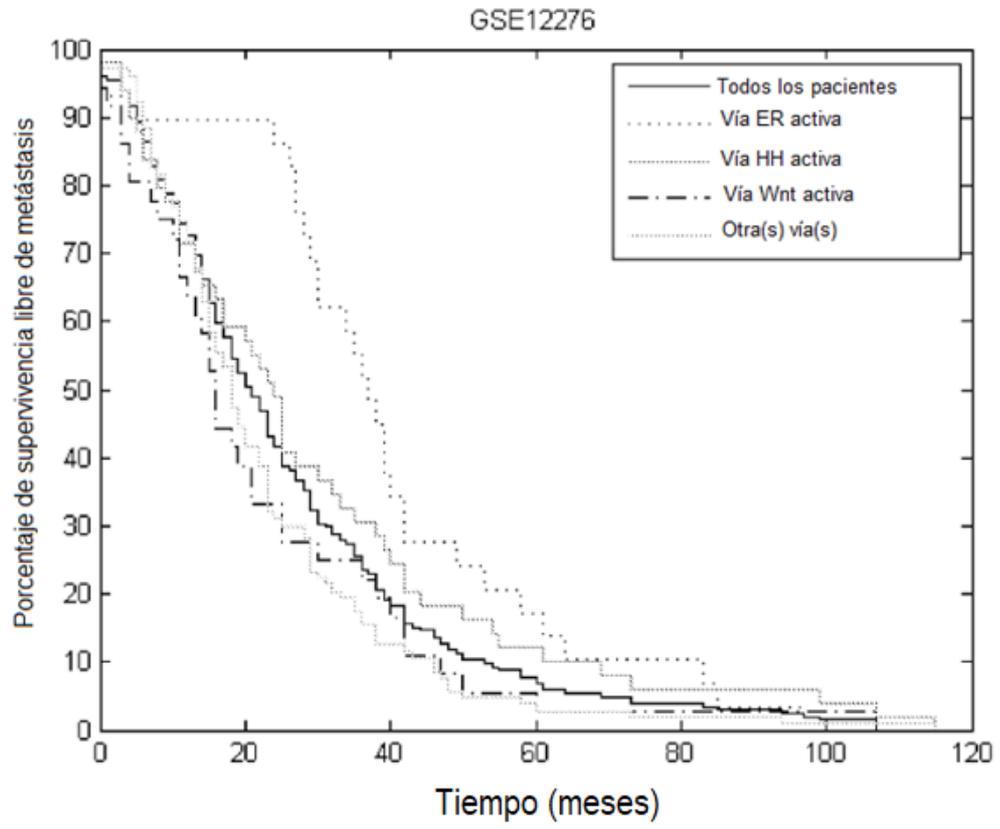


FIG. 38

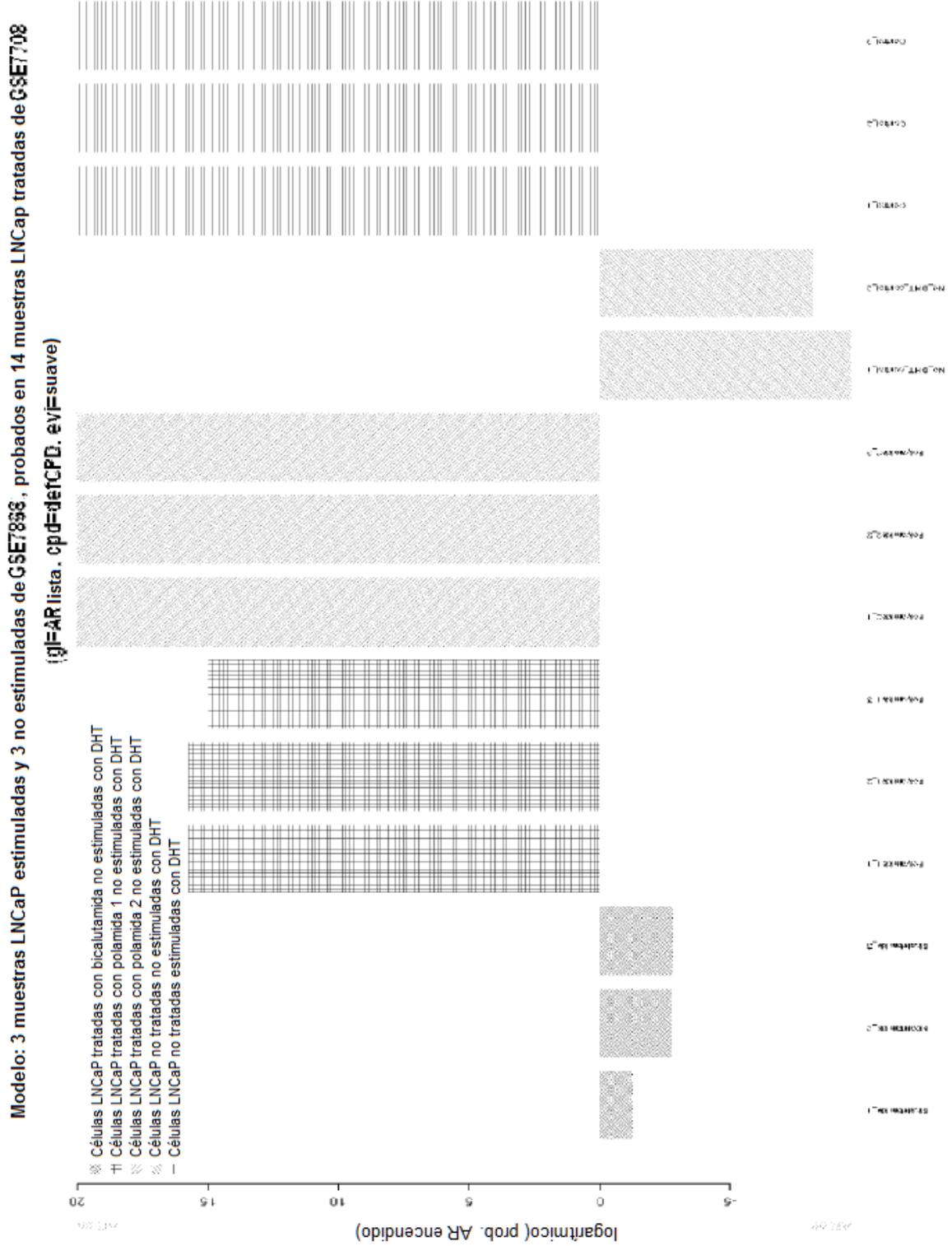


FIG. 39

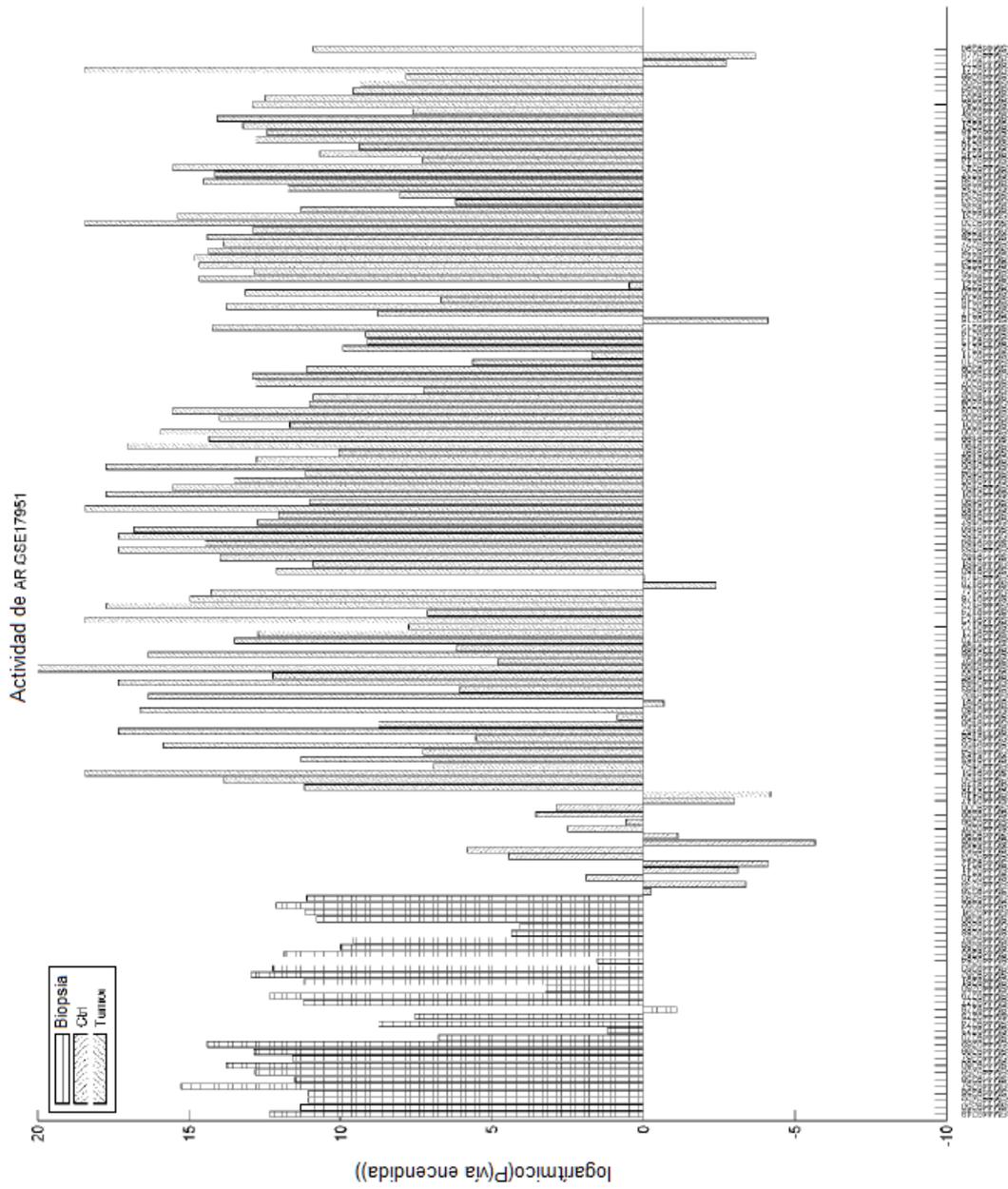


FIG. 40

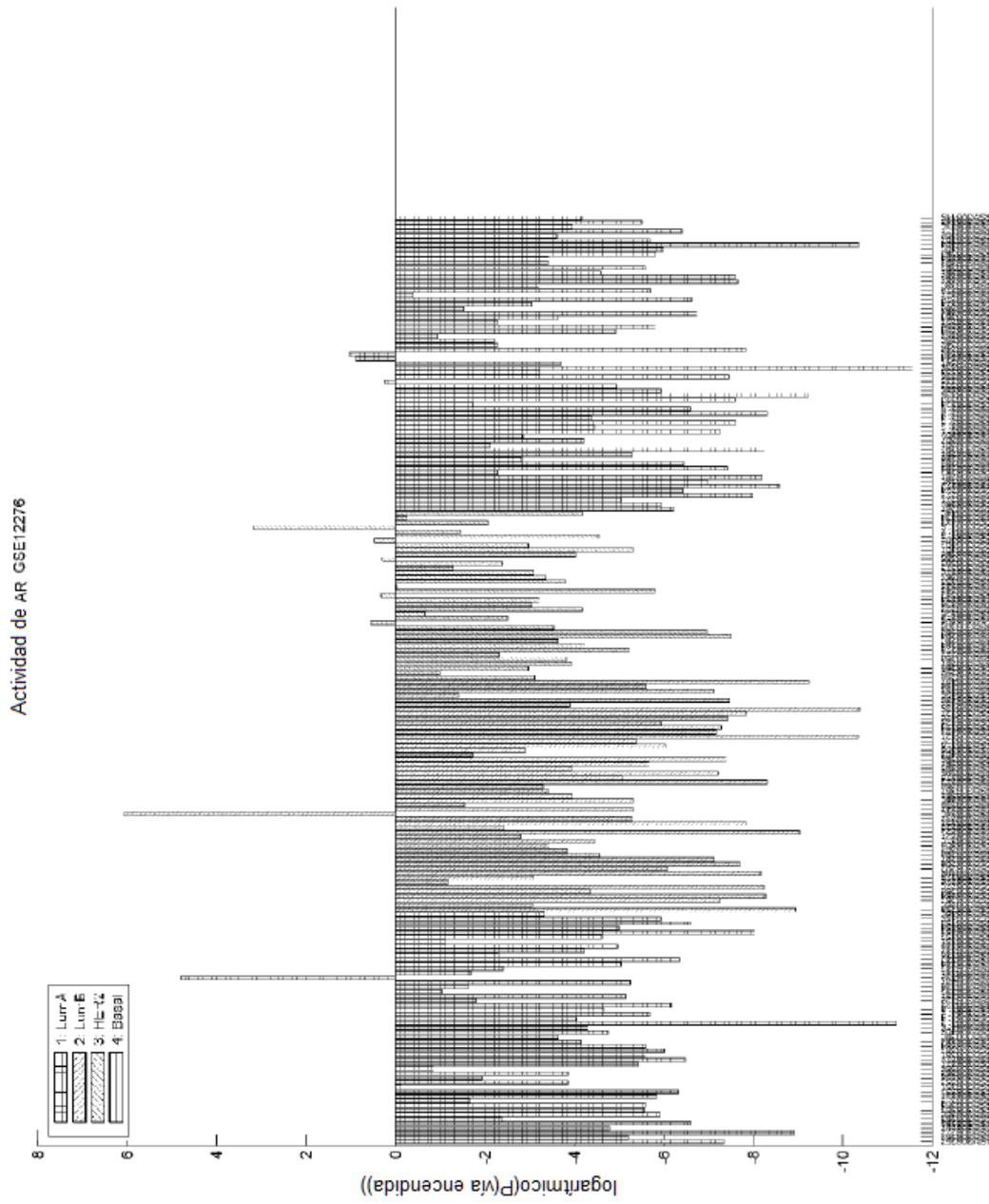


FIG. 41

Actividad de AR 55E36133

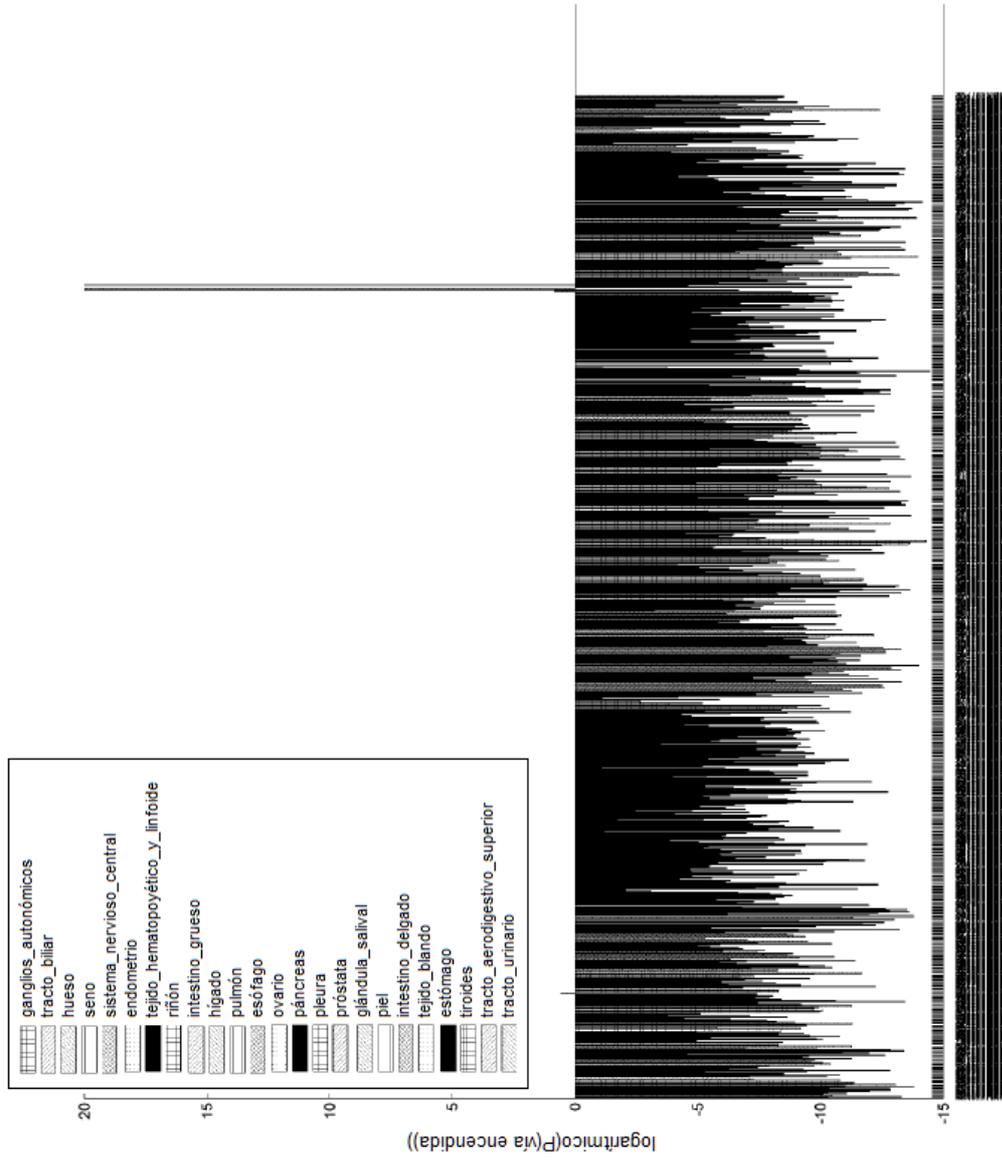


FIG. 42

