(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **3 003 262**

(51) Int. Cl.:

C12Q 1/6809 (2008.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **21.08.2020 PCT/GB2020/052018**

(87) Fecha y número de publicación internacional: **25.02.2021 WO21032996**

(96) Fecha de presentación y número de la solicitud europea: **21.08.2020 E 20761882 (8)**

(97) Fecha y número de publicación de la concesión europea: **11.12.2024 EP 4017997**

(54) Título: **Procedimiento de haplotipaje**

(30) Prioridad:

22.08.2019 GB 201912103

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
10.03.2025

(73) Titular/es:

OXFORD UNIVERSITY INNOVATION LIMITED (100.00%)
Buxton Court, 3 West Way
Oxford, OX2 0JB, GB

(72) Inventor/es:

HUGHES, JAMES R.

(74) Agente/Representante:

PONTI & PARTNERS, S.L.P.

DESCRIPCIÓN

Procedimiento de haplotipaje

[0001] La presente invención se refiere a un procedimiento para identificar mutaciones en alelos de un gen que pueden ser causa de desregulación de los niveles de expresión de los alelos del gen en una célula eucariota diana. La divulgación también se refiere a la detección de la expresión aberrante de genes que pueden estar asociados con una enfermedad o trastorno utilizando la fase de haplotipos. En particular, la divulgación también se refiere a un procedimiento para obtener una indicación de desregulación entre los niveles de expresión de al menos dos alelos de un gen en una célula eucariota diana.

[0002] Los estudios de asociación de genoma completo (GWAS, del inglés “Genoma-wide association studies”) de variantes comunes han identificado decenas de miles de variantes genéticas asociadas con rasgos complejos (por ejemplo, índice de masa corporal, altura, etc.) y predisposición a enfermedades comunes (por ejemplo, diabetes, enfermedades cardíacas, trastornos inmunológicos y predisposición al cáncer). A pesar de la genética estadísticamente robusta, la identificación de las variantes causales y los genes subyacentes a las señales de GWAS se ha visto frustrada por dos amplios factores. En primer lugar, las secuencias causales han sido complejas de localizar de manera precisa debido al extenso vínculo de variación de secuencia dentro del genoma; y en segundo lugar, independientemente del GWAS particular en cuestión, la gran mayoría de las variantes se encuentran fuera del genoma codificante interpretable.

[0003] La capacidad de interpretar el genoma no codificante es una prioridad para las ciencias biomédicas. En la última década, nuestra comprensión del ADN no codificante ha evolucionado desde "basura" hasta el complejo cableado regulador que controla la expresión y la función del transcriptoma codificante y no codificante. En el centro de esta función se encuentran los elementos reguladores no codificantes: promotores, potenciadores y límites intrón/exón. Aunque la existencia de algunos elementos reguladores se conoce desde hace varias décadas, su ubicuidad general (con funciones clave específicas de tejido en el desarrollo y los procesos celulares rutinarios) solo ha resultada clara en las últimas dos décadas. La gran superposición entre estos elementos y la genética de enfermedades comunes ha comenzado a dar sentido a la confusa distribución no codificante de las señales de GWAS.

[0004] En particular, se necesitan estudios adicionales del transcriptoma (es decir, el conjunto de todas las moléculas de ARN en una célula o una población de células) para ayudar a identificar genes que se expresan de manera aberrante en enfermedades específicas (o sus rutas bioquímicas asociadas) y, por lo tanto, para identificar dianas farmacológicas candidatas.

[0005] La expresión aberrante de genes puede deberse a cambios en los elementos que regulan la expresión de los genes. Los cambios en estos elementos pueden conducir a la sobreexpresión o subexpresión de los genes, a cambios en la expresión temporal de los genes o en su especificidad tisular.

[0006] Se sabe que dichos cambios también pueden conducir a una desregulación en los niveles de expresión de diferentes alelos del mismo gen dentro de una sola célula o población de células; y que dicha desregulación también puede estar asociada con enfermedades o trastornos específicos.

[0007] La determinación del vínculo entre la variación de secuencia a través de los alelos dentro de una célula (es decir, el haplotipo) se conoce como "fase"; y un hallazgo de diferentes niveles de expresión de los alelos se conoce como "sesgo alélico" (“allelic skew”, en inglés).

[0008] Tradicionalmente, la mayoría de los análisis bioquímicos y genéticos del ARNm se realizan en ARNm poliA$^+$. El ARN total se extrae primero de las células; a continuación se pasa por una columna de oligo-dT para unirse al ARNm poliA$^+$; y a continuación el ARNm poliA$^+$ se eluye selectivamente de la columna. El ARN poliA$^+$ contiene la secuencia codificante del ARNm, es decir, se han eliminado los intrones (no codificantes), y esta es la forma de ARNm que suele ser de mayor interés para el investigador. El uso de una columna de poli-dT también tiene la ventaja de que el ARNr (que es un contaminante no deseado y que se expresa abundantemente en todas las células) se elimina mediante este procedimiento.

[0009] Los inventores han reconocido ahora que se puede obtener información adicional del transcriptoma utilizando ARNm que no se limita a ARN poliA$^+$. Más específicamente, han reconocido que, si se utiliza pre-ARNm (por ejemplo, ARNm poliA$^+$, que todavía contiene intrones), entonces las variaciones naturales en las secuencias del genoma que se encuentran en los intrones y secuencias en dirección 3’ y que se eliminan en ARN poliA$^+$ se pueden utilizar para determinar el sesgo alélico y así identificar genes desregulados. Es importante destacar que, cuando se realiza en una secuencia genómica en fases, todas las variantes de secuencia se pueden atribuir a un alelo específico para determinar el sesgo alélico en todo el gen y vincularlas con la variación de secuencia fuera del cuerpo del gen.

[0010] Si bien el aislamiento de ARNm de poliA$^-$ es previamente conocido (por ejemplo, Kowalczyk, 2012), no se ha utilizado anteriormente en el contexto de haplotipaje genético o sesgo alélico.

[0011] El documento EP 1829979 A1 se refiere a procedimientos para buscar polimorfismos genéticos (por ejemplo, SNP) en ADNc como un medio para identificar genes cuyos niveles de expresión son diferentes entre alelos. Los inventores reconocieron que el ARNm poli A+ maduro tenía solo una secuencia de exones después del empalme y, por lo tanto, dichas secuencias eran "demasiado cortas" para comprender suficientes SNP para ser evaluadas. Por lo tanto, los inventores seleccionaron ARN intranuclear para proporcionar una "cadena larga" que se esperaba que contuviera muchos polimorfismos genéticos que pudieran permitir distinguir un gen cuyo nivel de expresión es diferente entre alelos (véanse los párrafos [0006] y [0018]).

[0012] James *et al.* (2013) utiliza pre-ARN para detectar sesgos en la expresión de un alelo en relación con otro en condiciones estimuladas en células inmunes, pero no lo hace en el contexto de un genoma en fase para vincularlo con la genética o para probar sesgos funcionales en el epigenoma dentro o distal a un gen.

[0013] Sigurdsson *et al.* (2008) utilizaron el sesgo alélico en el contexto de un genoma no en fase para detectar el sesgo alélico con un haplotipo de riesgo que se encuentra dentro del cuerpo del gen, pero no lo hacen en el contexto de un genoma en fase para vincularlo con la genética o para probar el sesgo funcional en el epigenoma dentro o distal a un gen.

[0014] Thomas *et al.* (2011) combina los usos del sesgo epigenético y el sesgo de expresión génica para observar los efectos epigenéticos asociados con un gen que es monoalélico, pero solo dentro del cuerpo del gen, ya que no utilizan esto en el contexto de un genoma en fases.

[0015] Rainbow *et al.* (2008) utilizan una forma de pre-ARNm para observar las diferencias específicas de alelos en la expresión del gen de IL-2, pero esto no está en el contexto de un genoma en fase para vincularlo con el panorama regulatorio distal o la genética.

[0016] La invención se describe en el conjunto de reivindicaciones adjuntas. En particular, la presente invención proporciona un procedimiento para identificar mutaciones en alelos de un gen que pueden ser causa de desregulación de los niveles de expresión de los alelos del gen en una célula eucariota diana.

[0017] La divulgación también se refiere a un procedimiento para obtener una indicación de desregulación entre los niveles de expresión de diferentes alelos del mismo gen en una célula eucariota diana, comprendiendo el procedimiento el uso de pre-ARNm. En un genoma en fases, esta desregulación puede vincularse entonces a una variación de secuencia fuera del cuerpo del gen que causa la desregulación del gen en el alelo afectado.

[0018] Sin la fase, cada SNP detectado en el análisis de ARN (a menos que estén cerca uno del otro, es decir, dentro de una distancia corta <300 pb) no se puede asignar al mismo alelo sin la suposición de que todos los SNP están sesgados en la misma dirección. Por lo tanto, no se pueden utilizar para agregar robustez estadística al sesgo alélico de un alelo determinado. Con la fase, todos los SNP se pueden asignar a un alelo determinado y se puede observar estadísticamente que se comportan de manera similar, en términos de cambios en la representación en el ARN y la dirección del sesgo.

[0019] Más importante aún, sin la fase, esta expresión de ARN sesgada no se puede vincular a cambios fuera del gen. Los elementos reguladores que controlan la expresión génica en *cis* en el mismo cromosoma pueden estar situados hasta a 2 millones de pares de bases del gen que controlan. Mediante la fase, los genotipos que vinculan genéticamente un cambio de secuencia con un rasgo o enfermedad determinados pueden verse como si estuvieran en el mismo alelo que un gen que se comporta de manera reproducible en su sesgo alélico en la expresión. Esto también permite vincular cambios epigenéticos en el SNP distal con cambios de expresión del alelo en el mismo cromosoma. Estos cambios epigenéticos se pueden detectar utilizando un sesgo alélico para ensayos genómicos basados en secuencias que miden la actividad potenciadora, tales como ensayos de cromatina abierta (por ejemplo, DNAse-seq o ATAC-seq), marcas de cromatina asociadas con la actividad reguladora (por ejemplo, ChIP-seq para marcas de cromatina, tales como H3k27ac, H3k4me1 o para la unión de proteínas, tales como la ARN polimerasa, la unión de factores de transcripción) o la unión de proteínas estructurales reguladoras importantes, tales como CTCF.

[0020] La presente divulgación se refiere a la detección de la expresión aberrante de genes que pueden estar asociados con una enfermedad o trastorno utilizando la fase de haplotipos.

[0021] Utilizando el procedimiento de la presente invención, resulta posible vincular el sesgo en la expresión génica con señales genéticas asociadas con rasgos y rasgos patológicos fuera del cuerpo del gen, donde generalmente se enriquecen. También permite vincular el sesgo en la expresión génica con el sesgo en ensayos genómicos basados en secuencias vinculados a actividades reguladoras, y puede hacerse de forma masiva y a escala del genoma, para validar el mecanismo subyacente a los cambios en la expresión génica.

[0022] La información que se obtiene a partir de los procedimientos de la presente invención se puede utilizar para identificar genes que se expresan de manera aberrante en enfermedades o trastornos específicos (o sus rutas bioquímicas asociadas) y, por lo tanto, para identificar dianas farmacológicas candidatas.

[0023] Esta información también puede ser útil para ayudar a determinar la causa genética subyacente de la enfermedad o trastorno.

[0024] La divulgación se refiere a un procedimiento para obtener una indicación de desregulación entre los niveles de expresión de al menos dos alelos del mismo gen en una célula eucariota diana, comprendiendo el procedimiento las etapas de:
para una pluralidad de genes de una o más células eucariotas diana,
(a) obtener pre-ARNm de al menos dos alelos del mismo gen; y
(b) determinar las relaciones ($R_{i,j}$) entre las cantidades de pre-ARNm de uno o más pares de alelos (i,j) del mismo gen;
en el que si $R_{i,j} \neq 1$ para uno o más pares de alelos (i,j) del mismo gen, entonces esto es indicativo de desregulación entre los niveles de expresión de esos dos alelos de ese gen en esa célula eucariota diana.

[0025] Preferiblemente, si $R_{i,j}$ es $< 0,9$ o si $R_{i,j} > 1,1$ para uno o más pares de alelos (i,j) del mismo gen, entonces esto es indicativo de desregulación entre los niveles de expresión de esos dos alelos de ese gen en esa célula eucariota diana.

[0026] La divulgación se refiere a un procedimiento para identificar genes cuyos alelos están desregulados en una célula eucariota diana, comprendiendo el procedimiento las etapas de:
para una pluralidad de genes de una o más células eucariotas diana,
(a) obtener pre-ARNm de al menos dos alelos de los genes; y
(b) determinar las relaciones ($R_{i,j}$) entre las cantidades de pre-ARNm de pares de alelos (i,j) de los genes;
en el que si $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen, entonces esto es indicativo de un gen cuyos alelos están desregulados en la célula eucariota diana.

[0027] Preferiblemente, si $R_{i,j}$ es $< 0,9$ o si $R_{i,j} > 1,1$ para un par de alelos (i,j) de un gen, entonces esto es indicativo de un gen cuyos alelos están desregulados en la célula eucariota diana.

[0028] La divulgación se refiere a un procedimiento para obtener una indicación de la causa de una enfermedad o trastorno, comprendiendo el procedimiento las etapas de:
para una pluralidad de genes de una o más células eucariotas diana,
(a) obtener pre-ARNm de al menos dos alelos de los genes; y
(b) determinar las relaciones ($R_{i,j}$) entre las cantidades de pre-ARNm de pares de alelos (i,j) de los genes;
en el que las células eucariotas diana son aquellas que están asociadas con una enfermedad o trastorno o son características de una enfermedad o trastorno,
y en el que si $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen, entonces esto es indicativo de que la enfermedad o trastorno es causado por una desregulación entre los niveles de expresión de los alelos de ese gen.

[0029] Preferiblemente, si $R_{i,j}$ es $< 0,9$ o si $R_{i,j} > 1,1$ para un par de alelos (i,j) de un gen, entonces esto es indicativo de que la enfermedad o trastorno es causado por una desregulación entre los niveles de expresión de los alelos de ese gen.

[0030] La divulgación se refiere a un procedimiento para identificar mutaciones en alelos de un gen que pueden ser causantes de una desregulación de los niveles de expresión de los alelos del gen en una célula eucariota diana, comprendiendo el procedimiento las etapas de:
para una pluralidad de genes de una o más células eucariotas diana,
(a) obtener pre-ARNm de al menos dos alelos de los genes; y
(b) determinar las relaciones ($R_{i,j}$) entre las cantidades de pre-ARNm de uno o más pares de alelos (i,j) de los genes;
en el que cuando $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen,
o en respuesta a la determinación de que $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen,
el procedimiento comprende además las etapas:
(c) determinar las secuencias de nucleótidos de ese par de alelos; y
(d) comparar las secuencias de nucleótidos de ese par de alelos con el fin de identificar diferencias entre las secuencias de nucleótidos de ese par de alelos;
en el que una o más de las diferencias entre las secuencias de nucleótidos del par de alelos del gen pueden ser mutaciones que son causantes de la desregulación de los niveles de expresión de los dos alelos de ese gen en la célula eucariota diana.

[0031] Preferiblemente, cuando $R_{i,j}$ es $< 0,9$ o si $R_{i,j} > 1,1$ para un par de alelos (i,j) de un gen, o en respuesta a la determinación de que $R_{i,j}$ es $< 0,9$ o si $R_{i,j} > 1,1$ para un par de alelos (i,j) de un gen, el procedimiento comprende las etapas:
(c) determinar las secuencias de nucleótidos de ese par de alelos; y
(d) comparar las secuencias de nucleótidos de ese par de alelos con el fin de identificar diferencias entre las secuencias de nucleótidos de ese par de alelos;
en el que una o más de las diferencias entre las secuencias de nucleótidos del par de alelos del gen pueden ser

mutaciones que son causantes de la desregulación de los niveles de expresión de los dos alelos de ese gen en la célula eucariota diana.

[0032] Tal como se utiliza en el presente documento, el término "desregulación" se refiere a diferencias en la regulación entre uno o más alelos de un gen. Normalmente, los alelos de un gen se expresan a niveles similares. Sin embargo, las mutaciones en las regiones reguladoras que controlan la expresión de los alelos individuales pueden dar como resultado una expresión mejorada o reducida de esos alelos. Por lo tanto, el término "desregulación" también se refiere a diferencias en los niveles de expresión de los alelos de un gen.

[0033] Las células diana son células eucariotas (es decir, células cuyo ADN nuclear tiene intrones). Preferiblemente, las células eucariotas son células de mamíferos, por ejemplo, células humanas, de mono, de ratón, de rata, de cerdo, de cabra, de caballo, de oveja o de vaca. Lo más preferible es que las células sean células humanas.

[0034] Las células pueden ser células primarias o células de una línea celular. Preferiblemente, las células son células primarias.

[0035] En algunas realizaciones de la presente invención, las células diana son células hematopoyéticas, por ejemplo, eritrocitos, linfocitos (por ejemplo, células T, células B y células asesinas naturales), granulocitos, megacariocitos y macrófagos. Preferiblemente, las células diana son células linfoides primarias.

[0036] En otras realizaciones, las células diana son células cerebrales; preferiblemente las células diana son células neuronales primarias.

[0037] Aunque las células son preferiblemente células diploides, los procedimientos de la invención son aplicables a células de otras ploidías, por ejemplo, células tetraploides.

[0038] En algunas realizaciones de la presente invención, las células eucariotas diana son aquellas que están asociadas con o son características de una enfermedad o trastorno. A continuación, se dan ejemplos de células eucariotas que están asociadas con o son características de una enfermedad o trastorno:

Células eucariotas	Trastorno o enfermedad
células cancerosas	Cáncer
Células neuronales	Demencia, enfermedad de Alzheimer y trastornos psiquiátricos (por ejemplo, esquizofrenia) y trastornos del espectro autista
Células pancreáticas	Diabetes
Células inmunes	Enfermedades autoinmunes e infecciones
Células eritroides	Malaria
Células del músculo cardíaco	Arritmias cardíacas, miocardiopatías
Células pulmonares	Enfermedad pulmonar, incluido el asma
Células de la piel	Afecciones dermatológicas, tales como la psoriasis y el eczema
Adipocitos	Obesidad
Retina	Degeneración de la retina
Células del tracto gastrointestinal superior e inferior	Enfermedad inflamatoria intestinal, síndrome del intestino irritable
Células hepáticas	Diabetes, esteatohepatitis no alcohólica, enfermedades del hígado
Células endoteliales y otros tipos de células vasculares	Enfermedad cardiovascular, incluidos accidentes cerebrovasculares y cardiopatía isquémica

[0039] Los ejemplos de enfermedades y trastornos incluyen aquellos en la tabla anterior.

[0040] El pre-ARNm se obtiene de una o más células eucariotas diana. En algunas realizaciones preferidas, el pre-ARN se obtiene de células diana individuales. En otras realizaciones, el ARN se obtiene de una población de células diana. En dichas realizaciones, la población de células diana comprende sustancial o completamente el mismo tipo de células, es decir, la población es sustancial o completamente homogénea.

[0041] Los genes son aquellos que están presentes en la célula eucariota diana. Tal como se utiliza en el presente

documento, el término "gen" no se limita a la secuencia codificante de ARN o de proteína. Incluye elementos reguladores asociados, por ejemplo, potenciadores, promotores y secuencias terminadoras. El término "gen" puede definirse como que incluye sus regiones transcritas, su intrón, regiones promotoras y todos los elementos reguladores o estructurales que determinan su actividad o nivel de expresión dentro de los 2 millones de pares de bases de los promotores de la parte transcrita del gen.

[0042] Los procedimientos de la presente invención son particularmente aplicables a genes que se ven afectados por una variación genética que se sabe que está asociada con una enfermedad o rasgo. La variación genética generalmente se habrá identificado mediante estudios de asociación de todo el genoma (GWAS), pero la variación genética también podría ser una mutación en un sujeto individual o en un subclón canceroso.

[0043] Aunque la presente invención es capaz de detectar la desregulación en cualquier gen, se prefieren ciertas clases de genes desregulados. Estos incluyen genes que se expresan específicamente en el tipo celular de interés en comparación con los genes expresados en todos los tipos celulares. De manera similar, también se prefieren los genes que son importantes para la función de un tipo celular de interés, por ejemplo, los genes de la sinapsis inmunitaria en las células inmunitarias. También se prefieren los genes que codifican factores de transcripción que determinan la identidad del tipo celular y/o el comportamiento de un tipo celular determinado. Además, se prefieren los genes que afectan la aptitud general de la célula, tal como los genes reguladores del ciclo celular o los genes relacionados con la apoptosis. En las células cancerosas, se prefieren los genes que afectan a procesos, tales como la proliferación, la movilidad y/o la adhesión celular, incluidos, pero sin limitarse a los mismos, los receptores de membrana, las moléculas mensajeras secundarias, los factores de transcripción y los reguladores epigenéticos.

[0044] Tal como se utiliza en este documento, el término "pluralidad de genes" se refiere a 2 o más genes, por ejemplo, 2-10, 10-100, 100-500, 500-1000, 1000-5000, 5000-10000 o 10000 o más.

[0045] Hay al menos dos alelos del mismo gen en cada célula eucariota diana. Por ejemplo, puede haber 2, 3 o 4 alelos del mismo gen. Preferiblemente, hay 2 alelos del mismo gen en cada célula eucariota diana.

[0046] Las secuencias de uno o más o todos los exones en los alelos de pre-ARNm pueden ser iguales o diferentes. Las secuencias de uno o más de todos los intrones en los alelos de pre-ARNm pueden ser iguales o diferentes.

[0047] Las células diana son preferiblemente heterocigóticas para uno o más de los alelos de interés.

[0048] La etapa (a) del procedimiento de la presente invención comprende la etapa de obtención de pre-ARNm. Preferiblemente, los pre-ARNm proceden de al menos dos alelos del mismo gen en un haplotipo determinado.

[0049] El pre-ARNm es la primera forma de ARN que se crea a través de la transcripción en el proceso de síntesis de proteínas. El pre-ARNm se transcribe a partir de una plantilla de ADN en el núcleo de la célula diana. El pre-ARN generalmente tendrá una caperuza en 5' (es decir, con una 7-metilguanosina) porque esta caperuza se produce unos pocos nucleótidos después del inicio de la síntesis de ARN. En el contexto de la presente invención, el pre-ARNm puede o no comprender una caperuza en 5'.

[0050] Existen dos diferencias principales entre el pre-ARNm y el ARNm en las células eucariotas: (i) se añade una cola de poli-A al extremo 3' del pre-ARNm; y (ii) se eliminan los intrones del pre-ARNm. Una vez que el pre-ARNm se ha procesado para incluir las características anteriores, se lo denomina "ARN mensajero maduro" o simplemente "ARN mensajero" (ARNm). El ARNm poliadenilado también se conoce como "ARN poliA^+".

[0051] En el contexto de la presente invención, es importante que todos o sustancialmente todos los intrones se conserven en el pre-ARNm, es decir, que no se hayan eliminado todos. Tal como se utiliza en el presente documento, en una realización, el término "sustancialmente todos los intrones se conservan" se utiliza para significar que al menos el 50%, 60%, 70%, 80%, 90%, 95% o 99% de la secuencia de nucleótidos intrónicos (que normalmente se eliminaría) se conserva en el pre-ARNm. En otra realización, el término "sustancialmente todos los intrones se conservan" se utiliza para significar que al menos el 50%, 60%, 70%, 80%, 90%, 95% o 99% del número total de intrones se conserva en el pre-ARNm.

[0052] En el contexto de la presente invención, también es importante (pero no esencial) que no se haya producido poliadenilación. Esto permite obtener información del extremo 3' del gen diana. Por lo tanto, el pre-ARNm puede contener cantidades insustanciales o trazas de ARNm poliadenilado.

[0053] En algunas realizaciones de la presente invención, también es importante que se conozca la variación de secuencia en torno a cada alelo. Esto permite que la variación de secuencia que afecta a la regulación génica se vincule con el sesgo alélico y la desregulación génica.

[0054] En los procedimientos de la presente invención, las cantidades óptimas de pre-ARNm (es decir, la máxima posible) se obtienen a partir de células eucariotas diana.

[0055] El pre-ARNm se puede obtener a partir de células mediante lisis celular seguida de extracción con disolvente y agotamiento de especies de ARNr y ARN poliA$^+$. El pre-ARNm también se puede obtener a partir de células mediante lisis celular seguida de unión de ARN a una columna de afinidad y agotamiento de especies de ARNr y ARNr poliA$^+$.

[0056] Preferiblemente, el pre-ARNm se obtiene mediante lisis celular seguida de extracción con disolvente y precipitación del ARN total. Por ejemplo, el ARN total se puede extraer utilizando reactivo TRIzol (Sigma), tubos PhaseLock Gel (5Prime) y centrifugación. El ARN se puede precipitar mezclándolo con volúmenes iguales de 2-propanol seguido de centrifugación. El ARN precipitado se puede lavar utilizando etanol al 75 %, secar y disolver en agua (por ejemplo, Kowalczyk, 2012).

[0057] El pre-ARNm que se utiliza en los procedimientos de la presente invención también puede mezclarse con cantidades insignificantes de otro ARN, por ejemplo, ARNr y/o ARNt o ADN.

[0058] Preferiblemente, el pre-ARNm no comprende ARN ribosómico (ARNr) o el pre-ARNm ha sido desprovisto de ARNr, por ejemplo, el ARNr se elimina del pre-ARNm antes de su uso. Por ejemplo, el ARNr se puede eliminar utilizando el kit RiboMinus Eukaryote para secuenciación de ARN (Invitrogen).

[0059] Preferiblemente, el pre-ARNm es ARNm poliA$^-$. Preferiblemente, el pre-ARNm no comprende ADN; éste puede eliminarse con una ADNasa.

[0060] La etapa (b) de los procedimientos de la presente invención comprende determinar las relaciones ($R_{i,j}$) entre las cantidades de pre-ARNm de uno o más pares de alelos (i,j) de los (mismos) genes.

[0061] En este sentido, los procedimientos de la presente invención abarcan tanto (1) procedimientos que implican la determinación de las cantidades absolutas de pre-ARNm de pares de alelos (y la relación entre ellas) como (2) procedimientos que determinan las cantidades relativas de pre-ARNm de pares de alelos (sin determinar necesariamente las cantidades absolutas de pre-ARNm de los alelos).

[0062] Los procedimientos para obtener las cantidades absolutas de pre-ARNm de alelos incluyen RNA-Seq seguido de alineamiento del genoma y recuento del número de secuencias alineadas.

[0063] Las secuencias derivadas de alelos específicos se identifican y cuentan identificando, dentro de los datos de RNA-Seq, los cambios de secuencia en los intrones, exones y regiones transcritas en dirección 3', que se sabe que son específicos de ese alelo.

[0064] Preferiblemente, las cantidades absolutas de pre-ARNm del primer y segundo alelos se determinan mediante RNA-Seq específica de la cadena, seguida de alineamiento del genoma y recuento del número de secuencias alineadas. Las secuencias que derivan de la cadena no transcrita se descartan para eliminar la contaminación genómica y la transcripción antisentido. Las secuencias que derivan de alelos específicos se identifican y cuentan identificando, dentro de los datos de RNA-Seq, los cambios de secuencia en los intrones, exones y regiones transcritas en dirección 3', que se sabe que son específicos de ese alelo (por ejemplo, Quinn *et al.,* 2013).

[0065] Preferiblemente, las cantidades relativas o absolutas de pre-ARNm del primer y segundo alelos se determinan utilizando RNA-Seq.

[0066] Las diferencias relativas en las cantidades del primer y segundo alelos se pueden determinar utilizando la hibridación de ADNc con micromatrices de SNP y determinando la señal relativa de cada alelo. Las diferencias relativas en las cantidades del primer y segundo alelos también se pueden determinar utilizando PCR específica de SNP y determinando la señal relativa de cada alelo.

[0067] $R_{i,j}$ se define en el presente documento como la relación de la cantidad de pre-ARNm de un par de alelos (i,j) del gen, donde i es el primer alelo y j es el segundo alelo.

[0068] En la expresión normal, la cantidad de pre-ARNm del primer alelo (i) del gen debe ser aproximadamente igual a la cantidad de pre-ARNm del segundo alelo (j) del gen, es decir, ambos alelos deben expresarse por igual.

[0069] Cuando $R_{i,j}$ = t- 1 para uno o más pares de alelos (i,j) del mismo gen, entonces esto es indicativo de desregulación entre los niveles de expresión de esos dos alelos de ese gen en esa célula eucariota diana.

[0070] En una realización, el término "$R_{i,j} \neq 1$" significa que R_{ij} no es sustancialmente igual a 1, por ejemplo, R_{ij} no es igual a 1 cuando se tiene en cuenta la variación aleatoria y el error experimental.

[0071] En otra realización, el término "$R_{i,j} \neq 1$" significa que la cantidad de pre-ARNm del primer alelo (i) del gen es estadísticamente diferente de la cantidad de pre-ARNm del segundo alelo (j) del gen (por ejemplo, $p < 0,05$ utilizando la prueba t de Student).

[0072] Preferiblemente, el término "$R_{i,j} \neq 1$" significa que $R_{i,j}$ es menor que 0,95, 0,9, 0,8, 0,7, 0,6, 0,5, 0,4, 0,3, 0,2, 0,1, 0,05 o 0,01; o $R_{i,j}$ es mayor que 1,05, 1,1, 1,2, 1,4, 1,6, 2, 2,5, 3, 5, 10, 20 o 100. Más preferiblemente, $R_{i,j}$ es menor que 0,9 o $R_{i,j}$ es mayor que 1,1.

[0073] Por ejemplo, si $R_{i,j}$ es < 0,9, entonces esto es indicativo de una desregulación entre la expresión del primer y segundo alelo del gen. En este caso, el nivel de expresión del primer alelo del gen es menor que el nivel de expresión del segundo alelo del gen.

[0074] Por ejemplo, si $R_{i,j}$ > 1,1, entonces esto es indicativo de una desregulación entre la expresión del primer y segundo alelo del gen. En este caso, el nivel de expresión del primer alelo del gen es mayor que el nivel de expresión del segundo alelo del gen.

[0075] La cuantificación de R_{ij} entre un alelo afectado (i) y un alelo normal (j) permite la identificación reproducible y estadísticamente robusta de genes vinculados a enfermedades o rasgos humanos. El grado de cambio en R_{ij} proporciona la dirección de la genética, es decir, muestra si un aumento o una disminución en la actividad del gen está asociada con la enfermedad o el rasgo.

[0076] Además, el tamaño de R_{ij} muestra en qué grado es necesario cambiar la actividad del gen para tener un efecto fisiológico medible.

[0077] Si se encuentra una desregulación entre los pares de alelos (i,j) del gen, entonces se pueden determinar las secuencias de los pares de alelos para intentar establecer la razón de la desregulación, por ejemplo, utilizando RNA-Seq. De acuerdo con la presente invención, cuando $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen, o en respuesta a la determinación de que $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen, el procedimiento comprende adicionalmente las etapas de determinar las secuencias de nucleótidos de ese par de alelos.

[0078] RNA-Seq (secuenciación de ARN), también llamada secuenciación shotgun del transcriptoma completo (WTSS, del inglés “Whole transcriptome shotgun sequencing”), utiliza la secuenciación de próxima generación (NGS, del inglés “next-generation sequencing) para revelar la presencia y cantidad de ARN en una muestra biológica en un momento dado.

[0079] En el contexto de la presente invención, RNA-Seq comprende las siguientes etapas:
(i) El pre-ARNm se fragmenta *in vitro* y se copia en ds-ADNc (por ejemplo, utilizando transcriptasa inversa); y
(ii) A continuación, se secuencia el ds-ADNc, preferiblemente utilizando procedimientos de secuenciación de lectura corta y alto rendimiento (por ejemplo, NGS).

[0080] Estas secuencias pueden entonces alinearse con una secuencia genómica de referencia para reconstruir qué regiones genómicas se estaban transcribiendo. Estos datos pueden usarse para anotar dónde se encuentran los genes expresados, sus niveles de expresión relativos y cualquier variante de empalme alternativa.

[0081] Si los niveles de expresión de los dos alelos son diferentes (por ejemplo, $R_{i,j}$ < 0,9 o $R_{i,j}$ > 1,1), entonces esto proporciona una indicación de que existe un cambio en los elementos reguladores de un alelo que controla la expresión del gen. Los elementos reguladores pueden existir dentro de los intrones del gen o fuera del cuerpo del gen.

[0082] También se pueden utilizar procedimientos, tales como ATAC-seq y NG Capture-C, para identificar con más detalle la causa genética precisa de la desregulación alélica, particularmente en casos en los que la causa se debe a un cambio en los elementos reguladores.

[0083] El procedimiento de la presente invención también puede comprender la etapa de llevar a cabo un ensayo basado en secuencias (preferiblemente ATAC-seq, DNase-seq o ChIP-seq) que mide la actividad de los elementos reguladores para detectar el sesgo en el mismo alelo de los genes que se encuentran sesgados en RNA-seq. El procedimiento de la presente invención incluye el uso de la fase de haplotipos.

BREVE DESCRIPCIÓN DE LAS FIGURAS

[0084]
La Figura 1 muestra el uso de pre-ARNm para identificar la desregulación génica utilizando haplotipos en fase y pre-ARNm.

La Figura 2 muestra el uso de pre-ARNm en genomas en fase para detectar la desregulación del gen *IKZF1* asociado con una variante de secuencia específica en un elemento regulador.

La Figura 3 muestra la frecuencia con la que se encuentran heterocigotos informativos en la población general para los alelos de riesgo asociados con enfermedades comunes.

EJEMPLOS

[0085] La presente invención se ilustra adicionalmente mediante los siguientes ejemplos, en los que las partes y porcentajes se expresan en peso y los grados son Celsius, a menos que se indique lo contrario. Debe entenderse que estos ejemplos, si bien indican realizaciones preferidas de la invención, se proporcionan únicamente a modo de ilustración.

Ejemplo 1: Uso de pre-ARNm en genomas en fase para detectar la desregulación de genes asociada con un haplotipo específico

[0086] La Figura 1A muestra una representación esquemática de dos alelos del genoma, cada uno de los cuales contiene dos genes y un elemento regulador. Los cambios de secuencia que distinguen a los dos alelos se muestran como X (por ejemplo, polimorfismos de un solo nucleótido (SNP), pequeñas inserciones o deleciones). Los exones de los dos genes se muestran como recuadros y los elementos promotores de los dos genes se muestran como una línea vertical y una flecha horizontal asociadas con el primer exón de los genes. La posición del elemento regulador se muestra como un triángulo entre los dos genes.

[0087] En el Haplotipo A, este elemento regulador contiene un cambio de secuencia que altera su actividad (mostrado como un tono más claro). Las interacciones reguladoras entre el elemento regulador y los genes, mapeadas por procedimientos 3C, tales como Capture-C, se muestran como líneas arqueadas con flechas. Los cambios de secuencia que distinguen el alelo fuente de los pre-ARNm de ambos genes se encuentran dentro de las partes transcritas de los genes (por ejemplo, intrones, exones y regiones en dirección 3'). En este ejemplo, la desregulación génica es causada por un elemento regulador dañado, pero el mismo uso de cambios de secuencia en fases combinados con la secuenciación de pre-ARNm se puede utilizar para cualquier otro mecanismo (por ejemplo, ganancia de función causada por variación de secuencia o variación estructural a mayor escala).

La Figura 1B muestra, en un gen de ejemplo, la mayor cobertura en formas pre-ARNm de RNA-Seq que retienen las regiones intrónicas y en dirección 3' transcritas del gen y que aumentan la cantidad de variación de secuencia capaz de detectar la desregulación de los genes. Los exones del gen se muestran como líneas verticales de grosor variable, mientras que los intrones se muestran como una línea sombreada horizontal.

La Figura 1C muestra la pérdida de unión del factor de transcripción GATA1 (ChIP qPCR) en un elemento regulador eritroide que contiene un solo cambio de par de bases (rs10758656), editado de forma homocigótica en células Hudep2.

La Figura 1D muestra la pérdida de la señal de cromatina abierta en el mismo elemento regulador utilizando ATAC-seq en presencia homocigótica del cambio de secuencia rs10758656.

La Figura 1E muestra la interacción específica eritroide entre este elemento regulador en células de tipo salvaje con el promotor de *RCL1* y *JAK2.*

La Figura 1F muestra la pérdida de expresión de sólo el gen *JAK2* en células editadas de forma homocigótica para rs10758656.

La Figura 1G muestra el sesgo alélico en las células eritroides primarias hacia el alelo de tipo salvaje tanto para ATAC-seq en el elemento regulador (cuadrados) como en la expresión de pre-ARNm (círculos) del gen *JAK2* solo en individuos heterocigótics para rs10758656.

Ejemplo 2: Uso de pre-ARNm en genomas en fase para detectar la desregulación del gen *IKZF1* asociada con una variante de secuencia específica en un elemento regulador

[0088]
La Figura 2A muestra la identificación de un cambio de secuencia en un elemento regulador que daña el potencial de unión de un factor de transcripción utilizando el algoritmo Sasquatch (Schwessinger R. *et al.,* 2017).

La Figura 2B muestra que este elemento regulador interactúa específicamente con el promotor del gen *IKZF1* mediante NG Capture-C.

La Figura 2C muestra el sesgo alélico de la señal de cromatina abierta hacia el tipo salvaje (Hap-B) en células eritroides primarias en 3 individuos heterocigóticos para el cambio de secuencia dañino (Hap-A) según lo determinado por ATAC-seq. La señal correspondiente dentro de la misma muestra en el haplotipo B está vinculada a la señal en el haplotipo A con una línea de puntos.

La Figura 2D muestra la disminución acumulada del pre-ARNm del haplotipo A del gen *IKZF1* , sumada a todos los cambios de secuencia transcritos que distinguen los dos haplotipos. La señal correspondiente dentro de la misma muestra en el haplotipo B está vinculada a la señal en el haplotipo A con una línea de puntos.

Ejemplo 3: El uso del sesgo alélico en el pre-ARNm en genomas en fase permite analizar la variación regulatoria a una escala sin precedentes en células primarias

[0089] La Figura 3 muestra el número de individuos informativos esperados para una frecuencia del alelo menor (MAF) determinada de la variante de secuencia en un muestreo aleatorio de la población general. Las barras grises representan el número de individuos esperados que son heterocigóticos en una MAF determinada. La línea negra muestra la distribución promedio de frecuencias de alelos menores en un estudio típico de asociación del genoma completo (GWA) para enfermedades humanas (diabetes tipo 1, espondilitis anquilosante, rasgos eritroides y esclerosis múltiple, combinados). Esto muestra que, con una MAF de 0,3, esto proporcionaría más de 20 observaciones independientes de desregulación génica y cubriría más de la mitad de un estudio típico de GWA. De manera similar, para una MAF de 0,1, esto proporcionaría 5 o más observaciones independientes de desregulación génica y cubriría más del 90 % de un estudio típico de GWA.

REFERENCIAS

[0090]

James C et al., Cell, vol. 155, 2013, "Human SNP Links Differential Outcomes in Inflammatory and Infectious Disease to a FOX03-Regulated Pathway", páginas 57-69.

Kowalczyk, M.S. et al. Intragenic enhancers act as alternative promoters. Mol Cell 45, 447-58 (2012).

Quinn EM, et al. (2013) Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. PLoS ONE 8(3): e58815. https://doi.org/10.1371/journal.pone.0058815.

Rainbow et al., BIOCHEMICAL SOCIETY TRANSACTIONS, vol. 36, 2008, "Commonality in the genetic control of Type 1 diabetes in humans and NOD mice: variants of genes in the IL-2 pathway are associated with autoimmune diabetes in both species", página 312.

Schwessinger R, et al. (2017) Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res. Oct. 2017;27(10):1730-1742. PMCID: PMC5630036.

Sigurdsson et al., HUMAN MOLECULAR GENETICS, vol. 17, 2008, "A risk haplotype of STAT4 for systemic lupus erythematosus is over-expressed, correlates with anti-dsDNA and shows additive effects with two risk alleles of IRF5", páginas 2868-2876.

Thomas et al., EPIGENETICS & CHROMATIC, vol. 4, 2011, "Allele-specific transcriptional elongation regulates monoallelic expression of theIGF2BPI gene", página 14.

REIVINDICACIONES

1. Procedimiento para identificar mutaciones en alelos de un gen que pueden ser causa de desregulación de los niveles de expresión de los alelos del gen en una célula eucariota diana,
en el que las mutaciones se encuentran en elementos reguladores que controlan la expresión de alelos individuales del gen y
en el que los elementos reguladores se encuentran fuera del cuerpo del gen,
comprendiendo el procedimiento las etapas de:
para una pluralidad de genes de una o más células eucariotas diana,
(a) obtener pre-ARNm de al menos dos alelos de los genes; y
(b) determinar las proporciones ($R_{i,j}$) entre las cantidades de pre-ARNm de uno o más pares de alelos (i,j) de los genes;
en el que cuando $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen,
o en respuesta a la determinación de que $R_{i,j} \neq 1$ para un par de alelos (i,j) de un gen,
el procedimiento comprende adicionalmente las etapas:
(c) determinar las secuencias de nucleótidos de ese par de alelos; y
(d) comparar las secuencias de nucleótidos de ese par de alelos con el fin de identificar diferencias de secuencia entre las secuencias de nucleótidos de ese par de alelos;
(e) atribuir todas las diferencias de secuencia a un alelo específico para determinar el sesgo alélico en el gen completo;
en el que el procedimiento se realiza en una secuencia de genoma en fase, y
(f) vincular estas diferencias de secuencia con la variación de secuencia en un elemento regulador fuera del cuerpo del gen.

2. Procedimiento, según la reivindicación 1, en el que la etapa (c) se lleva a cabo utilizando RNA-Seq.

3. Procedimiento, según la reivindicación 2, en el que las secuencias derivadas de alelos específicos se identifican y cuentan identificando, dentro de los datos de RNA-Seq, cambios de secuencia en los intrones, exones y regiones transcritas en dirección 3', que se sabe que son específicos de ese alelo.

4. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que si $Ri,j < 0,9$ o $Ri,j > 1,1$, entonces esto proporciona una indicación de que existe un cambio en los elementos reguladores en un alelo que controla la expresión del gen.

5. Procedimiento, según la reivindicación 4, en el que los elementos reguladores son aquellos que existen fuera de la región codificante del gen.

6. Procedimiento, según la reivindicación 2 o 3, en el que el procedimiento comprende además la etapa adicional de llevar a cabo un ensayo basado en secuencia que mide la actividad de los elementos reguladores para detectar sesgos en el mismo alelo de los genes que se encuentran sesgados utilizando RNA-seq.

7. Procedimiento, según la reivindicación 6, en el que el ensayo basado en secuencia es ATAC-seq, DNase-seq o ChIP-seq.

8. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que las células eucariotas son células linfoides primarias humanas o células neuronales primarias.

9. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que la pluralidad de genes es de 2 a 10, de 10 a 100, de 100 a 500, de 500 a 1000, de 1000 a 5000, de 5000 a 10000 o de 10000 o más genes.

10. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que hay 2 alelos del mismo gen en cada célula eucariota diana.

11. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que el pre-ARNm es ARNm poliA$^-$.

12. Procedimiento, según la reivindicación 11, en el que el pre-ARNm es ARNm poliA- obtenido a partir de ARN celular total.

13. Procedimiento, según cualquiera de las reivindicaciones anteriores, en el que $R_{i,j} \neq 1$ significa que $R_{i,j}$ es menor que 0,95, 0,9, 0,8, 0,7, 0,6, 0,5, 0,4, 0,3, 0,2, 0,1, 0,05 o 0,01; o $R_{i,j}$ es mayor que 1,05, 1,1, 1,2, 1,4, 1,6, 2, 2,5, 3, 5, 10, 20 o 100.

Figura 1

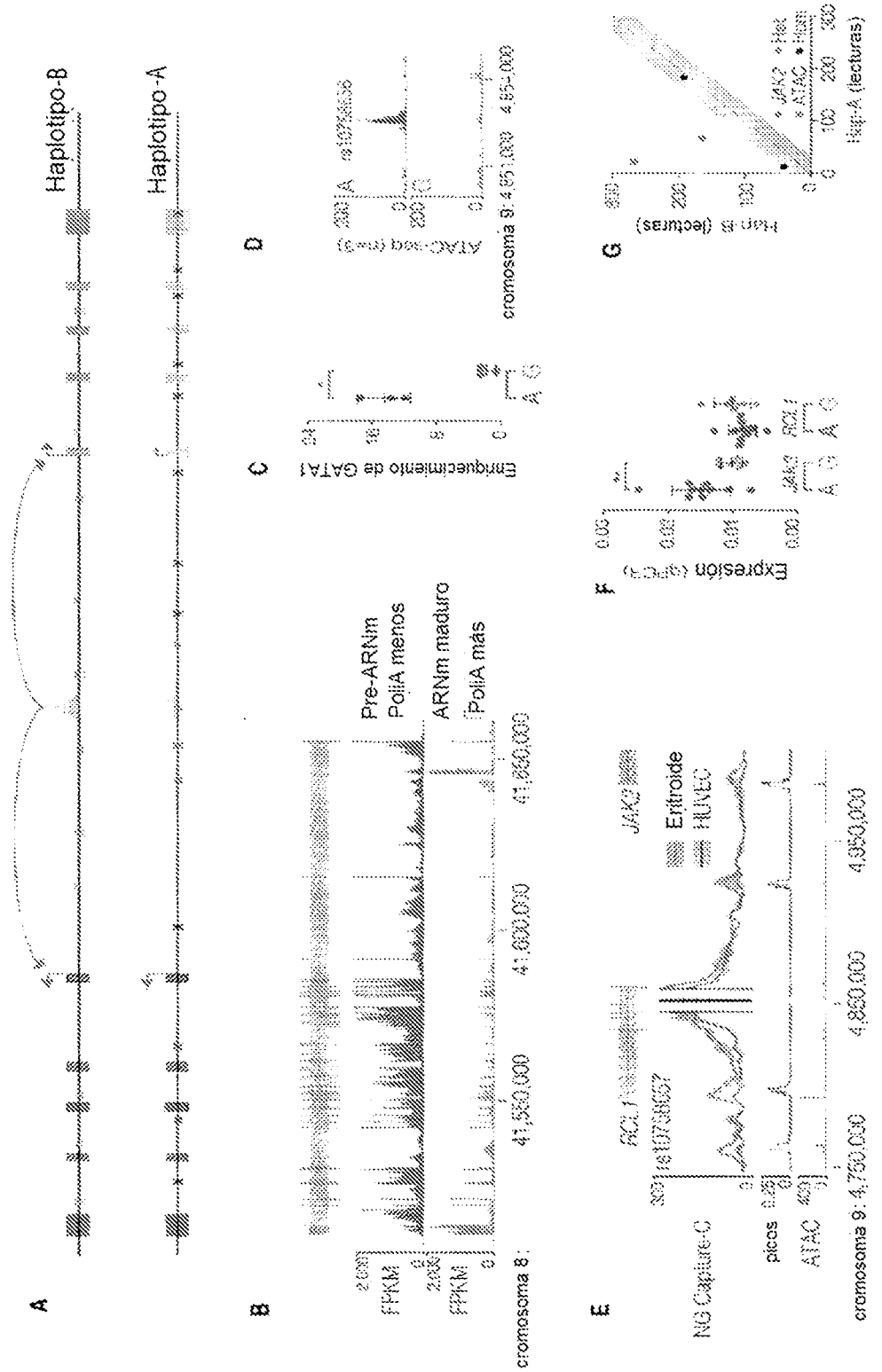

A
Haplotipo-B
Haplotipo-A
B
Pre-ARNm
PolíA menos
ARNm maduro
¡PolíA más
FPKM
cromosoma 8:
41,560,000
41,600,000
41,640,000
C
Enriquecimiento de GATA1
A
C
D
ATAC-seq (lecturas)
cromosoma 9: 4,851,000
4,85...000
E
JAK2
Eritroide
HUVEC
NG Capture-C
picos
ATAC
cromosoma 9: 4,750,000
4,800,000
4,850,000
F
Expresión (qPCR)
JAK2
ROLI
G
Hap-B (lecturas)
Hap-A (lecturas)
JAK2
ATAC
Het
Hom

Figura 2

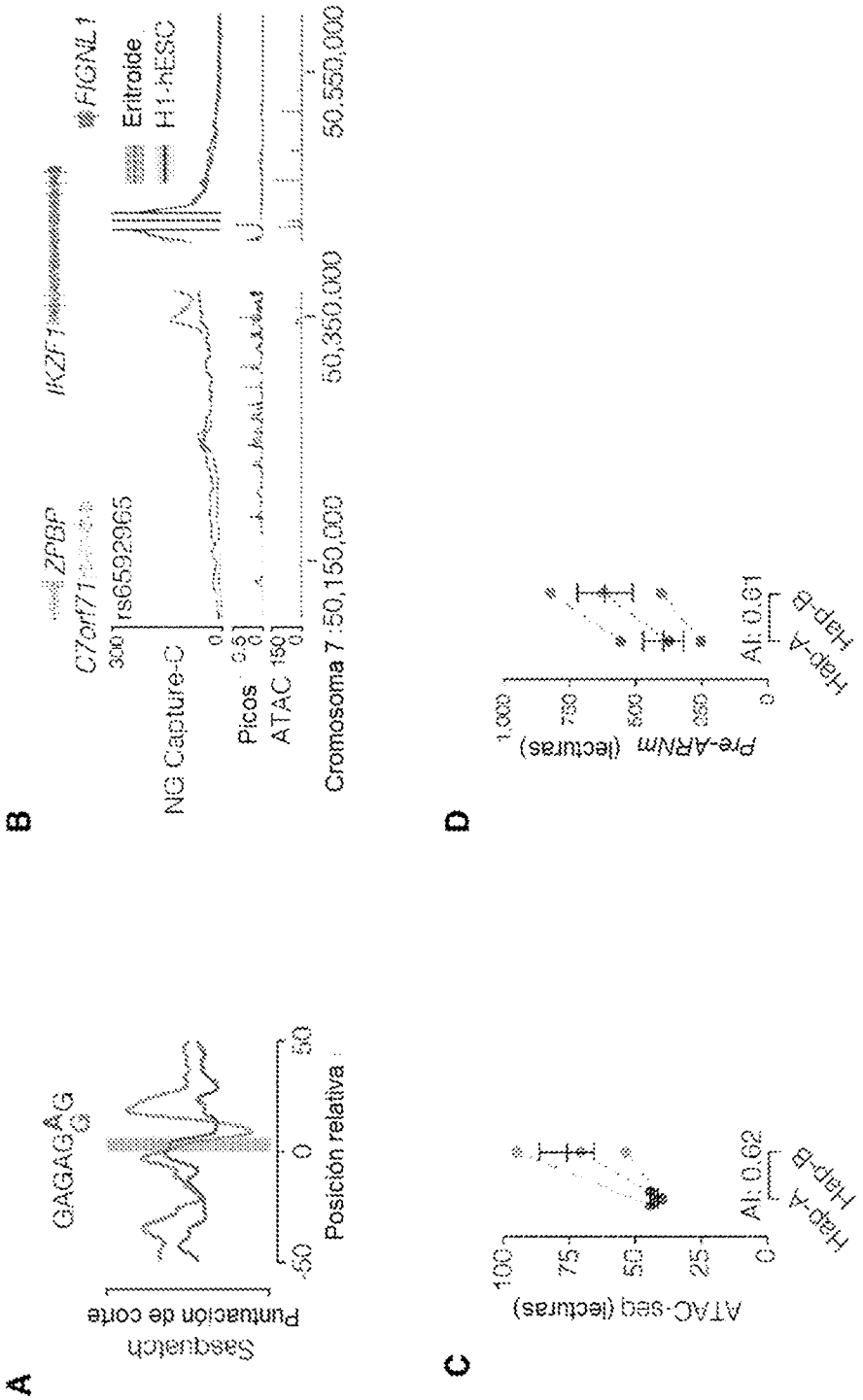
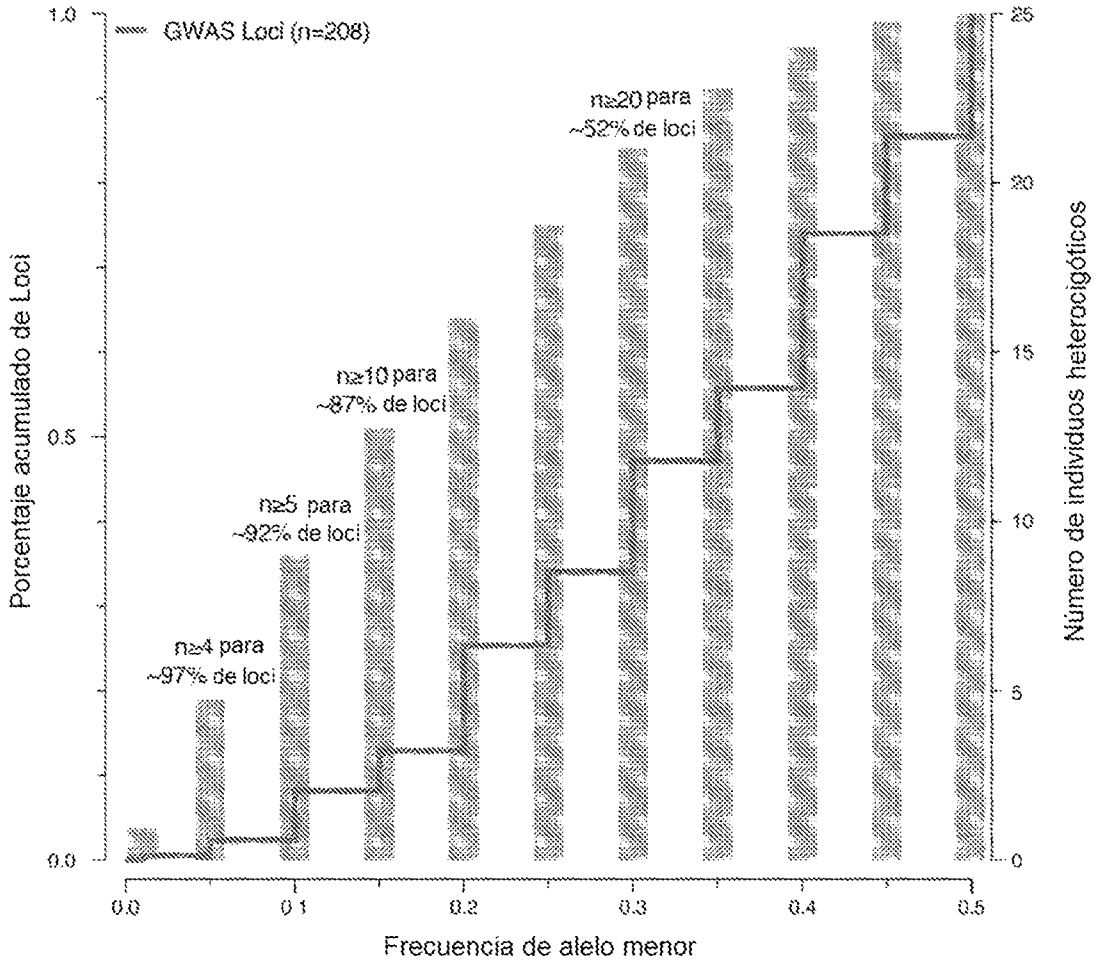

A
GAGAG
Secuenciación
Puntuación de corte
-50
0
50
Posición relativa
B
ZFBP
IKZF1
FIGNL1
Eritroide
H1-hESC
rs6592965
NG Capture-C
Picos
ATAC
Cromosoma 7: 50,150,000
50,350,000
50,550,000
C
ATAC-seq (lecturas)
100
75
50
25
0
AI: 0.62
Hap-A
Hap-B
D
Pre-ARNm (lecturas)
1,000
750
500
250
0
AI: 0.61
Hap-A
Hap-B

Figura 3


GWAS Loci (n=208)
n≥4 para ~97% de loci
n≥5 para ~92% de loci
n≥10 para ~87% de loci
n≥20 para ~52% de loci
Porcentaje acumulado de Loci
Número de individuos heterocigóticos
Frecuencia de alelo menor
0.0
0.5
1.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25