

19



OFICINA ESPAÑOLA DE
PATENTES Y MARCAS

ESPAÑA



11 Número de publicación: **2 964 691**

51 Int. Cl.:

G16B 20/00

(2009.01)

12

TRADUCCIÓN DE PATENTE EUROPEA

T3

86 Fecha de presentación y número de la solicitud internacional: **05.10.2022** **PCT/GB2022/052515**

87 Fecha y número de publicación internacional: **20.04.2023** **WO23062339**

96 Fecha de presentación y número de la solicitud europea: **05.10.2022** **E 22789995 (2)**

97 Fecha y número de publicación de la concesión europea: **18.10.2023** **EP 4200856**

54 Título: **Método implementado por ordenador y aparato para analizar datos genéticos**

30 Prioridad:

12.10.2021 GB 202114554

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

09.04.2024

73 Titular/es:

GENOMICS PLC (100.0%)
King Charles House, Park End Street
Oxford, Oxfordshire OX1 1JD, GB

72 Inventor/es:

WEALE, MICHAEL;
PLAGNOL, VINCENT YANN MARIE;
MOORE, RACHEL;
WELLS, DANIEL;
SETH, PRIYANKA y
PALMER, DUNCAN

74 Agente/Representante:

PONS ARIÑO, Ángel

ES 2 964 691 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

Método implementado por ordenador y aparato para analizar datos genéticos

- 5 La invención se refiere al análisis de datos genéticos y fenotípicos de un organismo para obtener información sobre el organismo, en particular para permitir estimaciones mejoradas del riesgo de que el organismo tenga un fenotipo objetivo o una combinación de fenotipos objetivo basándose en una ascendencia del organismo.

Antecedentes

- 10 Una puntuación de riesgo poligénico (PRS) es un resumen cuantitativo de la contribución del ADN heredado de un organismo a los fenotipos que puede exhibir. Una PRS puede incluir en su cálculo todas las variantes de ADN relevantes (ya sea directa o indirectamente) para un fenotipo de interés o puede usar sus componentes si son más relevantes para un aspecto particular de la biología de un organismo (incluidas células, tejidos u otras unidades, mecanismos o procesos biológicos). Una PRS puede usarse directamente, o como parte de una pluralidad de mediciones o registros sobre el organismo, para inferir aspectos de su biología pasada, actual y futura.

- 15 Las PRS están ganando terreno como herramienta para la prevención, estratificación y diagnóstico de enfermedades. En el contexto de la mejora de la salud humana y la asistencia sanitaria, las PRS tienen una variedad de usos prácticos, que incluyen, pero sin limitación: predecir el riesgo de desarrollar una enfermedad o fenotipo, predecir la edad de aparición de un fenotipo, predecir la gravedad de la enfermedad, predecir el subtipo de la enfermedad, predecir la respuesta al tratamiento, seleccionar estrategias de cribado apropiadas para un individuo, seleccionar intervenciones de medicación apropiadas y establecer probabilidades previas para otros algoritmos de predicción.

- 20 La PRS puede tener un uso directo como fuente de entrada en la aplicación de enfoques de inteligencia artificial y aprendizaje automático para realizar predicciones o clasificaciones a partir de otros datos de entrada de alta dimensión (por ejemplo, imagenología). Pueden usarse para ayudar a entrenar estos algoritmos, por ejemplo, para identificar mediciones predictivas basándose en datos no genéticos. Además de tener utilidad para hacer afirmaciones predictivas sobre un individuo, también se pueden usar para identificar cohortes de individuos, incluidas, pero sin limitación, las aplicaciones anteriores, calculando la PRS para un gran número de individuos y a continuación agrupando a los individuos sobre la base de las PRS.

- 25 Las PRS también pueden ayudar en la selección de personas para ensayos clínicos, por ejemplo, para optimizar el diseño del ensayo reclutando individuos con mayor probabilidad de desarrollar la enfermedad o los fenotipos relevantes, mejorando así la evaluación de la eficacia de un nuevo tratamiento. Las PRS contienen información sobre las personas para las que se calculan, pero también para sus familiares (que comparten una fracción de su ADN heredado). La información sobre el impacto del ADN de un individuo en sus fenotipos puede derivarse de cualquier evaluación relevante del impacto potencial de portar cualquier combinación particular de variantes de ADN.

- 30 A continuación se pone el foco en el análisis de la reciente riqueza de información que se deriva de los estudios de asociación genética (GAS). Estos estudios evalúan sistemáticamente la contribución potencial de las variantes de ADN a la base genética de un fenotipo. Desde mediados de la década de los 2000, se han llevado a cabo GAS (normalmente estudios de asociación de todo el genoma: GWAS, o estudios de asociación dirigidos a variantes únicas, o variantes en una región del genoma, o GWAS restringido a una región particular del genoma) en muchos miles de fenotipos (en su mayoría humanos), en millones de personas, generando miles de millones de vínculos potenciales entre genotipos y fenotipos. Los datos brutos resultantes a menudo se simplifican para producir datos estadísticos resumidos. Los datos estadísticos resumidos de GAS consisten en, para cada variante genética (ya sea imputada u observada), el tamaño del efecto inferido de la variante genética sobre el fenotipo del GAS y el error estándar del tamaño del efecto inferido. En otros casos, los datos a nivel individual, que consisten en un perfil genético completo de los individuos en un estudio e información sobre sus fenotipos, pueden estar disponible directamente. Sin embargo, los datos a nivel individual normalmente están menos disponibles debido a los requisitos sobre la privacidad de los datos de un individuo.

- 35 Las PRS agregan una gran cantidad de variantes genéticas ponderadas adecuadamente para proporcionar riesgos relativos específicos individuales de una enfermedad, tal como arteriopatía coronaria o cáncer de mama. Sin embargo, el mapeo de las variantes genéticas ponderadas al riesgo relativo no es sencillo. Una razón es el hecho de que las estrategias analíticas usadas para generar PRS generalmente arrojan puntuaciones no calibradas, es decir, la PRS no es fácilmente interpretable ni siquiera en la población para la que se obtuvo. Asimismo, la interpretación de la PRS varía entre poblaciones, porque el tamaño del efecto asociado con cada variante genética varía en función de la ascendencia. Por lo tanto, el "tamaño del efecto por unidad de PRS" agregado no es constante en todas las poblaciones humanas.

- 40 Dadas estas limitaciones, una opción práctica es probar una PRS en un conjunto de cohortes que representen diversas poblaciones humanas. A continuación se puede estimar el tamaño del efecto de la PRS en cada una de estas cohortes. Minxian Wang et al ("Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians", Journal of the American College of Cardiology, vol. 76, n.º 6, 3 de agosto de 2020, páginas 703-714) es un ejemplo de

este enfoque para una cohorte de ascendencia asiática del sur. Sin embargo, esto no siempre es posible porque los datos relevantes (que combinan datos genéticos y de resultados) pueden no existir, o puede que no todos los individuos encajen claramente en grupos de ascendencia bien definidos. Individuos de ascendencia mixta, o individuos provenientes de poblaciones pequeñas o poco estudiadas, a menudo no encajan en el pequeño número de grupos de ascendencia comúnmente usados.

El desafío es la extrapolación: basándose en casos y controles limitados o conjuntos de datos prospectivos de poblaciones particulares, es necesario extrapolar el tamaño del efecto de PRS a individuos que provienen de poblaciones diferentes o que no se ajustan exactamente a las características del conjunto de entrenamiento. Los métodos existentes para calcular la PRS y los riesgos relativos no tienen en cuenta adecuadamente estos factores, lo que lleva a puntuaciones que son inexactas para una gran proporción de individuos.

Para abordar estas y otras limitaciones, de acuerdo con un primer aspecto de la invención, se proporciona un método implementado por ordenador para analizar datos genéticos de acuerdo con la reivindicación 1.

El uso de la posición del individuo en un espacio de ascendencia para determinar la contribución genética al riesgo permite que el método tenga en cuenta a individuos que no encajan claramente en uno de un pequeño número de ascendencias predeterminadas. Esto puede mejorar la estimación del riesgo y la capacidad de realizar intervenciones apropiadas basándose en el riesgo genético.

La posición individual está representada por una combinación de variables continuas o pseudocontinuas, o la posición individual comprende una asignación a una combinación ponderada de una pluralidad de ascendencias. El uso de variables continuas o pseudocontinuas mejora aún más la capacidad de clasificar y comparar las posiciones. El uso de una combinación ponderada de ascendencias permite mejores predicciones para individuos de ascendencia mixta.

La contribución genética tiene una dependencia continua o pseudocontinua de la posición individual. Permitir que la contribución genética varíe al menos pseudocontinuamente con la posición mejora la resolución con la que a los individuos de ascendencia mixta se les puede asignar un riesgo apropiado.

En algunas realizaciones, la contribución genética comprende una suma de subcontribuciones para cada eje del espacio de ascendencia, cada subcontribución calculada usando una coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. Esto permite la mayor flexibilidad en cómo la contribución genética varía con la posición al tener en cuenta la posición relativa a cada eje del espacio de ascendencia.

En algunas realizaciones, el espacio de ascendencia es no isotrópico, de modo que la dependencia de cada subcontribución de la coordenada respectiva de la posición individual difiere entre las subcontribuciones. Esto mejora aún más la mayor flexibilidad en cómo la contribución genética varía con la posición en el espacio de ascendencia.

En algunas realizaciones, al menos dos de las subcontribuciones tienen dependencias de las coordenadas respectivas de la posición individual que están relacionadas por una probabilidad a priori compartida. Relacionar las dependencias en diferentes ejes introduce limitaciones que pueden reducir la tendencia al sobreajuste al determinar las dependencias.

En algunas realizaciones, la probabilidad a priori compartida se especifica de modo que las dependencias de las al menos dos subcontribuciones se muestrean a partir de la misma distribución. Esto puede resultar ventajoso cuando se sabe que la variación en diferentes ejes tendrá una forma funcional similar, pero puede variar en su valor específico.

En algunas realizaciones, la probabilidad a priori compartida se determina usando datos de entrenamiento de una pluralidad de individuos de entrenamiento y uno o más hiperparámetros predeterminados. El uso de hiperparámetros permite variar la fuerza de la limitación de ajuste dependiendo de la información disponible.

En algunas realizaciones, cada subcontribución comprende un producto de la puntuación de riesgo poligénico y la coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. Esta es una forma sencilla y eficaz de combinar la PRS y las coordenadas en el espacio de ascendencia.

En algunas realizaciones, calcular la contribución genética comprende calcular una distancia en el espacio de ascendencia entre la posición individual y una posición de referencia en el espacio de ascendencia, y calcular la contribución genética usando la distancia. Al reducir la variación para que dependa de una única distancia, se reduce el riesgo de extrapolar fuera de una región conocida del espacio de ascendencia.

En algunas realizaciones, la posición de referencia es una posición en el espacio de ascendencia de una ascendencia usada para entrenar coeficientes usados para calcular la puntuación de riesgo poligénico. Este es generalmente un punto de referencia apropiado, ya que representa el punto donde es probable que la PRS tenga el mayor tamaño del efecto, con una disminución probable en cualquier dirección.

En algunas realizaciones, el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad

de individuos de referencia que tienen una pluralidad de ascendencias diferentes; y calcular la contribución genética comprende ajustar a escala cada eje del espacio de ascendencia usando una varianza representada por el eje respectivo en los datos genéticos de referencia antes de calcular la distancia. Esta escala significa que los ejes más significativos en términos de la varianza que explican contribuirán más a la distancia, dándoles así mayor peso.

5 En algunas realizaciones, la distancia es una distancia euclidiana en el espacio de ascendencia. La distancia euclidiana es una medida de distancia sencilla y fácil de calcular.

10 En algunas realizaciones, la contribución genética comprende un producto de la puntuación de riesgo poligénico y la distancia. Esta es una forma sencilla y eficaz de combinar la PRS y la distancia en el espacio de ascendencia.

15 En algunas realizaciones, calcular la contribución genética comprende usar una dependencia lineal de la posición individual. Una dependencia lineal proporciona una manera sencilla de modelizar la dependencia con un comportamiento consistente.

20 En algunas realizaciones, calcular la contribución genética comprende usar una dependencia no lineal de la posición individual. Las dependencias no lineales permiten dependencias más complejas que pueden ser apropiadas en algunas circunstancias.

25 En algunas realizaciones, la dependencia no lineal comprende una función regularizada. El uso de una función regularizada garantiza que la dependencia tenga una forma razonable y fluida, y ayuda a evitar el sobreajuste, particularmente cuando los datos son escasos. Por ejemplo, en algunas realizaciones, la dependencia no lineal comprende una B-spline (línea polinómica suave básica) penalizada.

30 En algunas realizaciones, la dependencia no lineal se determina usando un proceso gaussiano como distribución previa para calcular la contribución genética mediante inferencia bayesiana. Los procesos gaussianos son adecuados para determinar dependencias no lineales, particularmente cuando los datos se obtienen de un proceso estocástico con una dependencia funcional desconocida, tal como para la genética.

35 En algunas realizaciones, el proceso gaussiano tiene un vector medio nulo. Esto simplifica el análisis del proceso gaussiano y no afecta a la generalidad del método, ya que los efectos medios se pueden añadir más adelante.

40 En algunas realizaciones, el proceso gaussiano tiene un vector medio correspondiente a una estimación previa de la contribución genética al riesgo para el individuo objetivo. Esto permite que el método controle el cociente de probabilidades en regiones del espacio de ascendencia alejadas de los datos genéticos de referencia, para representar el conocimiento sobre el riesgo en diferentes poblaciones.

45 En algunas realizaciones, una función núcleo del proceso gaussiano es una función estacionaria. Esta elección ayuda a garantizar que la variación de la contribución genética sea similar en diferentes partes del espacio de ascendencia, lo que generalmente se esperaría incluso si los valores absolutos difieren.

50 En algunas realizaciones, una función núcleo del proceso gaussiano decae a cero al disminuir la similitud entre muestras. Esto ayuda a garantizar que la contribución genética no se distorsione en regiones del espacio de ascendencia con datos genéticos de referencia relativamente escasos. Por ejemplo, en algunas realizaciones, la función núcleo es una función de base radial o una función de covarianza cuadrática racional.

55 En algunas realizaciones, se determina una distribución posterior para la inferencia bayesiana usando el proceso gaussiano y datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes. Esto vincula la distribución posterior al conjunto de datos particular en uso.

60 En algunas realizaciones, determinar la distribución posterior comprende aproximar la distribución posterior como una distribución gaussiana. Esto ayuda a mantener la manejabilidad, ya que algunas implementaciones pueden dar como resultado una distribución no gaussiana.

65 En algunas realizaciones, una función núcleo del proceso gaussiano depende de uno o más hiperparámetros. El uso de hiperparámetros para controlar la función núcleo protege contra el sobreajuste y refuerza un concepto significativo de distancia en el espacio de ascendencia. Por ejemplo, en algunas realizaciones, los hiperparámetros comprenden un hiperparámetro asociado con cada una de la puntuación de riesgo poligénico, la posición individual, y una interacción entre la puntuación de riesgo poligénico y la posición.

70 En algunas realizaciones, la contribución genética comprende un componente independiente de la ascendencia calculado usando la puntuación de riesgo poligénico que no depende de la posición individual. Esto permite que la PRS contribuya al riesgo hasta cierto punto, independientemente de la variación debida a la posición individual en el espacio de ascendencia.

75 En algunas realizaciones, la contribución genética comprende un componente dependiente de la ascendencia

calculado basándose en la posición individual que no depende de la puntuación de riesgo poligénico. Esto puede permitir que el riesgo tenga en cuenta un mayor riesgo debido a la ascendencia, independientemente de otras variaciones genéticas individuales.

5 El espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes. En algunas realizaciones, cada individuo de referencia se asigna a una de una pluralidad de ascendencias. El uso de datos genéticos de referencia de individuos que tienen diferentes ascendencias permite que el método tenga en cuenta una amplia gama de diferentes ascendencias para determinar el espacio de ascendencia y la posición individual.

10 En algunas realizaciones, un sistema de coordenadas del espacio de ascendencia se determina aplicando reducción de dimensiones a los datos genéticos de referencia. Esta es una técnica eficiente para caracterizar datos de alta dimensión, tales como datos genéticos, de una manera más compacta y eficiente. Por ejemplo, en algunas realizaciones, la reducción de dimensiones comprende el análisis de componentes principales o el análisis de componentes independientes.

15 En algunas realizaciones, la reducción de dimensiones comprende una discretización del espacio de ascendencia en un conjunto finito de ascendencias, y la posición individual comprende una proporción de pertenencia continua o pseudocontinua para cada ascendencia del conjunto finito de ascendencias. Esto proporciona una manera sencilla de expresar la posición individual en términos de los grupos de ascendencia definidos.

20 En algunas realizaciones, el sistema de coordenadas del espacio de ascendencia se elige para maximizar la varianza en los datos genéticos de referencia representados por el espacio de ascendencia. Esto asegura que la contribución genética tendrá en cuenta lo mejor posible la variación debida a la ascendencia.

25 En algunas realizaciones, el espacio de ascendencia tiene una dimensionalidad menor que la de los datos genéticos individuales, y determinar la posición individual comprende proyectar los datos genéticos individuales en el espacio de ascendencia. Esto permite calcular la contribución genética para un nuevo individuo.

30 En algunas realizaciones, se determina una dependencia de la contribución genética de la posición individual y la puntuación de riesgo poligénico usando datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes, comprendiendo los datos de entrenamiento, para cada uno de los individuos de entrenamiento, datos genéticos, y si el individuo de entrenamiento tiene el fenotipo o la combinación de fenotipos objetivo. Los datos de entrenamiento pueden ser o no los mismos que los datos genéticos de referencia, dependiendo de la información específica disponible de diferentes estudios.

35 En algunas realizaciones, los datos de entrenamiento comprenden además, para cada uno de los individuos de entrenamiento, datos informativos sobre una o más covariables no genéticas, y la contribución genética se estima conjuntamente en presencia de las covariables no genéticas; y el método comprende además recibir datos de covariables individuales para el individuo objetivo, los datos de covariables individuales informativos sobre las covariables no genéticas adicionales para el individuo objetivo. Esto permite que la contribución genética tenga en cuenta otros factores que pueden influir indirectamente en la contribución genética y que pueden estar correlacionados con la ascendencia. Por ejemplo, en algunas realizaciones, las covariables no genéticas incluyen una o más de peso, altura, características de comportamiento, rasgos médicos y otros biomarcadores, tales como mediciones basadas en sangre u orina.

En algunas realizaciones, el método comprende además generar la contribución genética al riesgo. Esto permite la aplicación posterior de la contribución genética.

50 En algunas realizaciones, el riesgo es un riesgo relativo en relación con un individuo que tiene una contribución genética estimada promedio, y el método comprende además: calcular el riesgo relativo para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la contribución genética y una contribución no genética al riesgo relativo; y generar el riesgo relativo. El riesgo relativo es una medida que puede correlacionarse más directamente con la probabilidad de que un individuo desarrolle un fenotipo.

55 En algunas realizaciones, calcular el riesgo relativo comprende usar una función de pérdida para determinar un valor del riesgo relativo para el individuo objetivo a partir de una distribución del riesgo relativo para el individuo objetivo. La función de pérdida determina una selección apropiada de un valor único como valor esperado a partir de la distribución. Por ejemplo, en algunas realizaciones, la función de pérdida es una función de error cuadrático medio o una función de pérdida exponencial asimétrica.

60 En algunas realizaciones, el método comprende además: calcular un riesgo absoluto para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la contribución genética calculada usando el método de cualquier reivindicación anterior; y generar el riesgo absoluto. Un riesgo absoluto es otra medida útil para expresar la probabilidad de que un individuo desarrolle un fenotipo particular.

De acuerdo con un segundo aspecto de la invención, se proporciona un aparato para analizar datos genéticos de acuerdo con la reivindicación 15. El procesador puede configurarse además para llevar a cabo operaciones análogas a las descritas anteriormente para el método implementado por ordenador.

- 5 La invención también puede materializarse en un programa informático que comprenda instrucciones que hagan que la ordenador lleve a cabo el método, o un medio legible por ordenador que comprenda instrucciones que, cuando son ejecutadas por un ordenador, hacen que el ordenador lleve a cabo el método.

10 Las realizaciones de la invención se describirán adicionalmente a modo de ejemplo solamente con referencia a los dibujos adjuntos, en los que:

La figura 1 muestra el cociente de probabilidades estimado por unidad de PRS de acuerdo con un método de la técnica anterior;

La figura 2 es un diagrama de flujo del método de acuerdo con una realización de la invención;

15 La figura 3 muestra datos de entrenamiento faltantes de individuos del sur de Asia proyectados en el espacio de ascendencia;

La figura 4 muestra estimaciones del tamaño del efecto para diferentes ascendencias bajo un esquema modelo de posición absoluta lineal de la contribución genética usando los datos de entrenamiento de la figura 3;

20 La figura 5 muestra estimaciones de máxima verosimilitud para diferentes términos de la contribución genética bajo el mismo modelo y condiciones que la figura 4

La figura 6 muestra estimaciones de tamaño del efecto similares a las de la figura 4, donde la contribución genética usa menos dimensiones del espacio de ascendencia;

25 La figura 7 muestra el cociente de probabilidades estimado por unidad de PRS para diferentes ascendencias bajo un modelo de posición relativa lineal de la contribución genética para los mismos datos de entrenamiento que las figuras 3-6;

La figura 8 muestra estimaciones de tamaño del efecto para diferentes ascendencias bajo el mismo modelo que la figura 7;

La figura 9 muestra estimaciones de tamaño del efecto para diferentes ascendencias bajo un modelo de posición absoluta lineal jerárquico de la contribución genética para los mismos datos de entrenamiento que las figuras 3-8;

30 La figura 10 muestra estimaciones de tamaño del efecto para diferentes ascendencias bajo un modelo de posición absoluta no lineal de la contribución genética para los mismos datos de entrenamiento que las figuras 3-9;

La figura 11 muestra estimaciones de tamaño del efecto para diferentes ascendencias bajo un modelo de posición relativa de proceso gaussiano de la contribución genética para los mismos datos de entrenamiento que las figuras 2-7.

Descripción detallada

Estado de la técnica

40 Como se ha mencionado anteriormente, el mapeo de variantes genéticas ponderadas en forma de una PRS al riesgo relativo de un individuo no es sencillo. El desafío es la extrapolación: basándose en casos y controles limitados o conjuntos de datos prospectivos de poblaciones particulares, es necesario extrapolar el tamaño del efecto de PRS a individuos que provienen de poblaciones diferentes o que no se ajustan exactamente a las características del conjunto de entrenamiento.

45 El enfoque más simple, es suponer que el tamaño del efecto de la puntuación de riesgo poligénico (PRS) estandarizada es constante en todas las poblaciones, o equivalentemente, que la varianza fenotípica explicada por la PRS estandarizada es independiente de la ascendencia. Esto es demostrablemente falso, pero puede proporcionar un punto de partida razonable cuando se dispone de datos limitados. Esto da como resultado el siguiente modelo:

$$Y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = \beta_0 + \beta_{PRS} X_{PRS,i} \quad (1)$$

55 dónde Y_i es una variable aleatoria que indica si el individuo i tiene una enfermedad (1) o no (0), π_i es la probabilidad de que el individuo i sea un caso, $X_{PRS,i}$ es la PRS del individuo i , β_{PRS} es el tamaño del efecto de la PRS, y β_0 es un factor constante. Se podrían incluir otras covariables, pero se han omitido aquí para mayor claridad. Algunos enfoques existentes incluyen componentes principales (PC, que se analizan con más detalle a continuación) o una proporción de ascendencia europea como covariables en este modelo, pero sin interacción con la PRS. Por lo tanto, esto todavía da como resultado un tamaño del efecto de PRS único. (Amariuta, T., Ishigaki, K., Sugishita, H. *et al.* 2020 Methods, Ec. 2, Fritsche *et al.* 2021 Methods, Ec. 1 y Bitarello y Mathieson 2021).

60 Un segundo enfoque, más razonable, que representa el estado actual de la técnica, consiste en asignar individuos a una de una colección de ascendencias predeterminadas y ajustar el modelo de la Ec. (1) en cada ascendencia. Esto da como resultado el siguiente modelo:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_{0,k} + \beta_{PRS_k} X_{PRS,i} \quad (2)$$

dónde β_{PRS_k} es el tamaño del efecto de PRS de la población k, y también se permite que el factor constante varíe entre poblaciones como $\beta_{0,k}$.

Este enfoque se visualiza en la figura 1. La figura 1 muestra muestras en un conjunto de datos de prueba proyectadas sobre los dos primeros componentes principales (PC) definidos por el conjunto de datos de 1.000 Genomas, y coloreados por el cociente de probabilidades (OR) estimado asociado a una desviación estándar de PRS. En el caso de los datos mostrados en la figura 1, el cociente de probabilidades por una desviación estándar de PRS simplemente corresponde a $e \beta_{PRS_k}$ en el modelo (2), porque la PRS está estandarizada para tener una distribución estándar de 1.

La figura 1 muestra una proyección de los individuos en un espacio de ascendencia usando los PC definidos por el conjunto de datos de 1000 Genomas. Este tipo de proyección en un espacio de ascendencia se analizará más adelante en relación con la presente invención, pero normalmente no se hace en los enfoques existentes. La proyección mostrada en la figura 1 sirve para demostrar los enfoques existentes en un formato que se compara fácilmente con realizaciones de la presente invención que se analizan más adelante.

Como se muestra, cada individuo es asignado estrictamente (o tipificado rigurosamente) a uno de los cinco grupos de ascendencia predefinidos, en concreto, europea (EUR), asiática del sur (SAS), nativa americana (AMR_NAT), asiática oriental (EAS) y africana (AFR_SS). En otras palabras, cada individuo es simplemente asignado al grupo de ascendencia al que es más similar en cierta medida. Los cocientes de probabilidades se estiman por separado para cada grupo tipificado rigurosamente, lo que da como resultado discontinuidades a lo largo de las líneas continuas de ascendencia entre europeos y asiáticos orientales y europeos y africanos. Existen discontinuidades similares a lo largo de las clinas hacia los asiáticos del sur y los nativos americanos a lo largo de los ejes definidos por los PC 3 y 4, aunque esto no se muestra aquí. Los puntos rojos indican centros de agrupamiento de muestras en cada una de las cinco superpoblaciones de 1.000 genomas, definidos por la media de las muestras asignadas a ese grupo después de la eliminación de las muestras mezcladas y la posterior asignación de agrupamientos.

Un estudio previo ha determinado distancias en un espacio de ascendencia para demostrar la atenuación del rendimiento predictivo, como se predijo por motivos teóricos (Priv6 *et al.* 2021). Sin embargo, el uso de distancias en un espacio de ascendencia no se ha relacionado con la importancia de los tamaños del efecto de PRS, ni se ha usado para corregir puntuaciones de riesgo basadas en la ascendencia. Algunos estudios han intentado cuantificar en qué medida la precisión de la PRS disminuye con la distancia genética (Priv6 *et al.* 2021), pero no ha habido ningún desarrollo de metodologías para dar cuenta de este fenómeno.

La estimación incorrecta del tamaño relativo del efecto crea conjuntos de datos no calibrados, que sobreestiman o subestiman el papel de la genética. Esto es potencialmente perjudicial ya que se pueden aplicar medidas preventivas o de diagnóstico a las personas equivocadas. Estimar el tamaño correcto del efecto de la PRS estandarizada requiere ponderar apropiadamente la genética de un individuo, donde la interpretación del tamaño del efecto en esta escala es el cambio en las probabilidades logarítmicas asociadas a un cambio de una desviación estándar en PRS. La presente invención aborda estos problemas. Para dejar de agrupar a todos los individuos en grupos homogéneos discretos, se permite que el tamaño del efecto de PRS varíe con la ubicación del individuo en el espacio de ascendencia y se construyen modelos que usan esta definición continua de ascendencia.

Introducción a la presente invención

La figura 2 muestra una realización de un método implementado por ordenador para analizar datos genéticos. El método comprende recibir S10 una puntuación de riesgo poligénico (PRS) 10 para un fenotipo objetivo o una combinación de fenotipos objetivo para un individuo objetivo. Como se ha mencionado anteriormente, una PRS es un resumen cuantitativo del efecto de las variantes genéticas de un individuo sobre su riesgo de un fenotipo o combinación de fenotipos particular. La PRS 10 puede calcularse usando cualquier método adecuado. El cálculo de la PRS 10 podrá realizarse inmediatamente antes del presente método, por ejemplo por el mismo sistema informático. Como alternativa, la PRS 10 puede calcularse en otro lugar en un momento diferente y transmitirse al sistema que lleva a cabo el presente método. La PRS 10 puede referirse a cualquier fenotipo. Por ejemplo, el fenotipo objetivo podría ser una enfermedad tal como una cardiopatía, cáncer, diabetes, o cualquier otra enfermedad de interés.

El método comprende además una etapa S20 de recibir datos genéticos individuales 20 para el individuo objetivo. Los datos genéticos individuales 20 son informativos sobre una ascendencia del individuo objetivo. Por ejemplo, los datos genéticos individuales 20 pueden comprender datos sobre una pluralidad de variantes genéticas que se sabe que son indicativas de la ascendencia de un individuo. Sin embargo, no es necesario que los datos genéticos individuales 20 incluyan información sobre variantes genéticas que sean informativas sobre el fenotipo objetivo o la combinación de fenotipos objetivo. La información sobre la genética del individuo que se relaciona directamente con el fenotipo objetivo ya está codificada en la PRS 10.

Espacio de ascendencia

El método comprende además determinar S30 una posición individual en un espacio de ascendencia usando los datos genéticos individuales. El espacio de ascendencia se puede definir usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes. Por ejemplo, los datos genéticos de referencia pueden derivarse de GWAS o de bases de datos disponibles públicamente, tales como la base de datos de 1000 Genomas mencionada en relación con la figura 1. Los datos genéticos de referencia son informativos sobre las ascendencias de los individuos de referencia. Por ejemplo, los datos genéticos de referencia pueden comprender datos para cada individuo de referencia sobre al menos una pluralidad de variantes genéticas que se sabe que son indicativas de la ascendencia de un individuo.

La pluralidad de variantes genéticas sobre las cuales los datos genéticos de referencia son informativos pueden ser iguales o superponerse sustancialmente con (por ejemplo, tener al menos el 50 % de las variantes iguales entre las dos pluralidades) la pluralidad de variantes genéticas sobre las cuales los datos genéticos individuales son informativos. Al usar las mismas variantes genéticas para definir el espacio de ascendencia que se usan para el individuo, se puede mejorar la precisión de la colocación del individuo en el espacio.

Cada individuo de referencia puede asignarse a una de una pluralidad de ascendencias. Por ejemplo, los individuos de referencia podrán ser asignados a uno de los grupos de ascendencia predefinidos mencionados anteriormente, en concreto, europea (EUR), asiática del sur (SAS), nativa americana (AMR_NAT), asiática oriental (EAS) y africana (AFR_SS). Se pueden incluir más o menos grupos de ascendencia predefinidos dependiendo de los datos genéticos de referencia que se usen. Etiquetar a los individuos de referencia de esta manera ayuda a definir las regiones en el espacio de ascendencia asociadas con cada grupo de ascendencia. Como alternativa, sólo un subconjunto de los individuos de referencia puede asignarse a una de la pluralidad de ascendencias. Esto puede ser preferible cuando los datos genéticos de referencia contienen individuos mezclados que no coinciden estrechamente con ninguno de los grupos de ascendencia predefinidos.

Es preferible incluir datos de tantos individuos de referencia como sea posible en los datos genéticos de referencia, y de individuos que tengan una amplia gama de ascendencias. Esto ayudará a definir mejor el espacio de ascendencia y permitirá que la interpolación sobre el individuo objetivo se realice con mayor confianza en una región más grande del espacio de ascendencia.

En algunas realizaciones, un sistema de coordenadas del espacio de ascendencia se determina aplicando reducción de dimensiones a los datos genéticos de referencia. En genética humana se sabe usar metodologías estadísticas de reducción de dimensiones para mapear el espacio genético altamente multidimensional (con millones de variantes) en un subespacio que tiene un número menor de dimensiones. Este subespacio es el espacio de ascendencia. La presente invención se refiere principalmente al uso de la posición en el espacio de ascendencia para corregir el tamaño del efecto de la PRS. Sin embargo, la elección del espacio de ascendencia y cómo se deriva tiene efectos posteriores sobre los modelos usados para estimar el tamaño del efecto de PRS.

En algunas realizaciones, la reducción de dimensiones puede comprender una discretización del espacio de ascendencia en un conjunto finito de ascendencias, y la posición individual comprende una proporción de pertenencia continua o pseudocontinua para cada ascendencia del conjunto finito de ascendencias. Por ejemplo, la posición individual puede comprender una asignación a uno o una combinación ponderada de una pluralidad de ascendencias. Al permitir que la posición individual indique una mezcla de diferentes ascendencias discretas para el individuo, se puede obtener una estimación más precisa de la ascendencia de un individuo en relación con el enfoque de "tipificación rigurosa" del estado de la técnica ilustrado en la figura 1. Esto, a su vez, permite una estimación más precisa del tamaño del efecto de la PRS para el individuo objetivo.

El espacio de ascendencia normalmente se aproxima mediante una reducción de dimensionalidad de los datos genéticos de referencia a k dimensiones. Normalmente se usa una técnica de reducción de dimensiones lineal. Por ejemplo, la reducción de dimensiones puede comprender análisis de componentes principales (PCA), análisis de componentes independientes, factorización matricial no negativa o análisis factorial. También se podrían usar técnicas de reducción de dimensiones no lineal siempre que sea posible proyectar nuevas muestras (es decir, los datos genéticos individuales del individuo objetivo) en el subespacio de dimensión reducida que es el espacio de ascendencia. Por ejemplo, el espacio de ascendencia se puede definir realizando un análisis de componentes principales (PCA) y tomando los primeros k componentes principales (PC). Normalmente, el espacio de ascendencia puede tener 2, 3 o 4 dimensiones. Sin embargo, también se pueden usar espacios de ascendencia de dimensiones superiores, por ejemplo que tienen 5, 6, 7, 8, 9, 10 o más de 10 dimensiones.

Para los ejemplos en el resto de esta solicitud, se usa el subespacio lineal abarcado por los primeros cuatro PC.

En algunas realizaciones, el sistema de coordenadas del espacio de ascendencia se elige para maximizar la varianza en los datos genéticos de referencia representados por el espacio de ascendencia. Por ejemplo, donde la reducción de dimensiones comprende PCA, esto se puede lograr seleccionando los PC que representan la mayor varianza en los datos genéticos de referencia.

Esta elección significa que los ejes del espacio de ascendencia corresponden mejor a variaciones que reflejan diferencias en la ascendencia. Por ejemplo, donde los datos genéticos de referencia comprenden datos sobre una pluralidad de variantes genéticas para cada individuo de referencia, la varianza en los datos genéticos de referencia puede ser la varianza entre variantes genéticas que se sabe que son indicativas de la ascendencia de los individuos, o entre variantes que se considera que varían entre grupos de ascendencia dentro del conjunto de datos.

En algunas realizaciones, el espacio de ascendencia tiene una dimensionalidad menor que la de los datos genéticos individuales, y determinar la posición individual comprende proyectar los datos genéticos individuales en el espacio de ascendencia. Una vez que se ha determinado el espacio de ascendencia, la posición individual se puede determinar proyectando los datos genéticos individuales en el espacio de ascendencia. La posición individual puede representarse de diversas maneras en el espacio de ascendencia. En algunas realizaciones, la posición individual puede representarse mediante una combinación de variables ordenables, variables pseudocontinuas o variables continuas. Por ejemplo, la posición individual puede comprender una variable ordenable, variable pseudocontinua, o variable continua para cada eje o dimensión del espacio de ascendencia. El tipo de variable puede ser el mismo para todas las dimensiones del espacio de ascendencia, o puede diferir entre las dimensiones del espacio de ascendencia.

A menudo resulta útil entender la PRS como el riesgo relativo condicionado a una posición en el espacio de ascendencia genéticamente definido. Una PRS se denomina "centrada" si la expectativa (media) de esa PRS condicionada a una posición en el espacio de ascendencia es 0. Una PRS (centrada) se denomina "estandarizada" si la varianza de la PRS condicionada a una posición en el espacio de ascendencia o etiqueta de población es 1. Se prefiere, aunque no es esencial, que se use una PRS centrada y estandarizada, de modo que la PRS refleja sólo el riesgo relativo para el individuo y no captura las diferencias entre los grupos de ascendencia.

Una PRS centrada y estandarizada puede verse como un gaussiano con media 0 y varianza 1 en cualquier punto del mapa de PC (es decir, el espacio de ascendencia), debido al teorema del límite central. Se pueden usar técnicas diferentes pero relacionadas para obtener esa estandarización, pero estas no son el foco de este documento.

Calcular la contribución genética

El método comprende además calcular S40 una contribución genética a un riesgo para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la puntuación de riesgo poligénico 10 y la posición individual.

Una vez determinada la posición individual en el espacio de ascendencia, se puede usar una relación entre la posición individual y el tamaño del efecto por unidad de PRS para determinar la contribución genética al riesgo para el individuo. La relación entre la posición individual, la PRS y la contribución genética al riesgo se pueden modelizar de diversas maneras. En el presente método, se consideran dos enfoques amplios. Como se ha mencionado anteriormente, los ejemplos en esta aplicación definen el espacio de ascendencia usando el subespacio lineal abarcado por los primeros cuatro PC definidos a partir de los datos del proyecto 1000 Genomas.

El primer enfoque amplio es un enfoque multidimensional de "posición absoluta". En dichas realizaciones, la contribución genética comprende una suma de subcontribuciones para cada eje del espacio de ascendencia, cada subcontribución calculada usando una coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. En otras palabras, las coordenadas de la posición individual en cada eje o dimensión del espacio de ascendencia contribuyen todas de forma independiente en la dependencia de la contribución genética de la posición individual. Para obtener la dependencia de la contribución genética de la posición individual y PRS, se ajusta un modelo que incluye la PRS y cada eje del espacio de ascendencia (definido por la reducción de dimensionalidad truncada de los datos genéticos de referencia), y sus interacciones, para obtener estimaciones específicas de ascendencia de la contribución del tamaño del efecto de PRS al riesgo.

En realizaciones de dicho enfoque de "posición absoluta", el espacio de ascendencia puede ser no isotrópico, de modo que la dependencia de cada subcontribución de la coordenada respectiva de la posición individual difiere entre las subcontribuciones. Esto permite la máxima flexibilidad para ajustar la dependencia de la contribución genética para dar cuenta de diferentes variaciones en la ascendencia del individuo objetivo, representada por cambios en la posición individual a lo largo de diferentes ejes del espacio de ascendencia.

El segundo enfoque amplio es un enfoque unidimensional de "posición relativa". En dicha realización, calcular la contribución genética comprende calcular una distancia en el espacio de ascendencia entre la posición individual y una posición de referencia en el espacio de ascendencia, y calcular la contribución genética usando la distancia. En otras palabras, la contribución genética depende sólo de la distancia absoluta en el espacio de ascendencia entre la posición individual y un punto de referencia. Por ejemplo, la distancia puede ser una distancia euclidiana en el espacio de ascendencia. La contribución genética puede comprender un producto de la puntuación de riesgo poligénico y la distancia.

En este conjunto de modelos de "posición relativa", la posición absoluta de cada muestra en el espacio de ascendencia no se incorpora directamente a la dependencia de la contribución genética, sino más bien se usan las distancias

relativas entre la posición individual del individuo objetivo y algún punto de referencia, como alguna función del eje del espacio de ascendencia multidimensional. En alguna realización, la posición de referencia es una posición en el espacio de ascendencia de una ascendencia usada para entrenar coeficientes usados para calcular la puntuación de riesgo poligénico. En otras palabras, la posición de referencia corresponde a la ascendencia de los datos de entrenamiento usados para crear la PRS, es decir, entrenar (por ejemplo, mediante un algoritmo de aprendizaje automático) un conjunto de coeficientes que se usan para obtener la PRS para el individuo objetivo basándose en sus datos genéticos individuales.

Este enfoque de "posición relativa" está motivado por la observación de que la varianza explicada por la PRS decae monótonamente con la distancia ancestral de la población usada para ajustar la PRS. Por tanto, esta distancia relativa y su interacción con PRS se incluyen como covariables al ajustar el modelo de dependencia de la contribución genética. Esto permite estimar los tamaños del efecto de la PRS que varían continuamente con la distancia ancestral de la población usada para entrenar la PRS.

En algunas realizaciones, el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes; y calcular la contribución genética comprende ajustar a escala cada eje del espacio de ascendencia usando una varianza representada por el eje respectivo en los datos genéticos de referencia antes de calcular la distancia. Cuando se usa el enfoque de posición relativa, ajustar a escala los ejes del espacio de ascendencia por la varianza que explican garantiza que las coordenadas que son más significativas tengan más peso en la posición relativa.

Ya sea en el primer enfoque (multidimensional) o en el segundo (unidimensional), la contribución genética puede tener una dependencia continua o pseudocontinua de la posición individual. Además, para cada uno de los enfoques multidimensionales y unidimensionales, se consideran dependencias tanto lineales como no lineales de la contribución genética en PRS y la posición individual. De manera específica, se consideran estimaciones basadas en regresión logística, que implícitamente suponen una dependencia lineal entre la posición individual y el cambio en el logaritmo del cociente de probabilidades, y estimaciones basadas no lineales de PRS específicas de ascendencia usando procesos gaussianos y modelos aditivos generales (GAM).

Modelos lineales

En algunas realizaciones, calcular la contribución genética comprende usar una dependencia lineal de la posición individual. Como se ha mencionado anteriormente, esto puede incorporar la posición individual como una posición absoluta multidimensional o una posición relativa unidimensional.

Posición absoluta

En algunas realizaciones, la contribución genética comprende una suma de subcontribuciones para cada eje del espacio de ascendencia, cada subcontribución calculada usando una coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. Un método sencillo para incorporar una estimación de PRS del tamaño del efecto dependiente de la ascendencia es incluir la PRS del individuo, la posición individual (en términos de los componentes principales en este ejemplo) y sus interacciones como covariables en el siguiente modelo logístico.

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_{PRS} X_{PRS,i} + \sum_{j=1}^k \left(\beta_{PC_j} X_{PC_j,i} + \beta_{PRS \times PC_j} X_{PRS,i} X_{PC_j,i} \right) \quad (3)$$

donde, como se ha indicado anteriormente, Y_i es una variable aleatoria que indica si el individuo i tiene una enfermedad (1) o no (0), π_i es la probabilidad de que el individuo i sea un caso, $X_{PRS,i}$ es la PRS del individuo i , β_{PRS} es el tamaño del efecto de la PRS sola, y β_0 es un factor constante. Además, $X_{PC_j,i}$ es la coordenada de la posición individual del individuo i para la dimensión (eje) del espacio de ascendencia correspondiente al PC j , β_{PC_j} es el tamaño del efecto para la posición sola, y $\beta_{PRS \times PC_j}$ es el tamaño del efecto para la combinación de posición y PRS. Cada subcontribución comprende un producto de la puntuación de riesgo poligénico y la coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia.

En este modelo, se supone algún tamaño del efecto general en ausencia de información ancestral, que a continuación es modificado por la posición individual en el espacio de ascendencia. Dado que PRS está estandarizada para tener media 0 y varianza 1, independientes de la ubicación ancestral, a continuación se puede evaluar el tamaño del efecto de PRS específico de la ascendencia $\beta_{PRS,i}$ para el individuo i , que es el tamaño del efecto para una desviación estándar de PRS:

$$\beta_{PRS,i} := \beta_{PRS} + \sum_{j=1}^k \beta_{PRS \times PC_j} X_{PC_j,i}$$

Esto simplifica la ecuación para el modelo (3) anterior a

$$\text{logit}(\pi_i) = \beta_0 + \beta_{PRS,i} X_{PRS,i} + \sum_{j=1}^k (\beta_{PC_j} X_{PC_j,i})$$

Dada la ubicación de los centros de agrupamiento en el espacio de PC, se puede usar esta fórmula para compararla con el enfoque categórico del estado de la técnica proporcionado por el modelo (2) anterior. Usando el modelo (3), también es sencillo evaluar los intervalos de confianza para las estimaciones del tamaño del efecto de la PRS en el presente modelo usando la estadística de Wald.

$$L = [0, 1, 0_{1 \times k}, X_{PC_1,i}, \dots, X_{PC_k,i}]$$

$$SE = \sqrt{LVL^T}$$

donde V es la matriz de covarianza. Entonces,

$$CI_{95\%} = \beta_{PRS,i} \pm 1.96 \times SE$$

El modelo (3) es simple, extremadamente rápido y proporciona una extensión continua o pseudocontinua al método más moderno de asignación discreta a categorías ancestrales. Como el modelo subyacente del riesgo latente es lineal, la interpretación de una unidad de PRS estandarizada no varía con la ubicación en la escala de PRS.

En esta y otras realizaciones, se determina una dependencia de la contribución genética de la posición individual y la puntuación de riesgo poligénico usando datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes. Los datos de entrenamiento comprenden, para cada uno de los individuos de entrenamiento, datos genéticos, y si el individuo de entrenamiento tiene el fenotipo o la combinación de fenotipos objetivo.

Vale la pena señalar que los datos de entrenamiento usados para entrenar el modelo de dependencia de la contribución genética de la PRS y la posición individual pueden diferir de los datos genéticos de referencia usados para definir el espacio de ascendencia. Por ejemplo, los datos genéticos de referencia pueden contener solo información sobre variantes genéticas relevantes para la determinación de la ascendencia genética, y pueden no contener suficientes datos relevantes para el fenotipo objetivo para permitir que también se usen para entrenar el modelo. En algunas realizaciones, los datos genéticos de referencia y los datos de entrenamiento pueden ser los mismos, pero esto no siempre es verdad.

La extrapolación a regiones del espacio de ascendencia no representadas en los datos de entrenamiento puede conducir a resultados falsos, por lo que se debe tener cuidado de interpolar y no extrapolar. Si una superpoblación (que corresponde sustancialmente a uno de los grupos de ascendencia predefinidos, por ejemplo) no está presente en los datos de entrenamiento pero sí en los datos genéticos de referencia, que se supone que capturan la diversidad genética mundial, entonces un solo PC distinguirá a esa superpoblación de las demás. Por otra parte, los individuos restantes de las otras superpoblaciones probablemente no diferirán en gran medida a lo largo del eje del espacio de ascendencia definido por ese PC. Esto da como resultado inestabilidad en la estimación del tamaño del efecto de ese único PC y vulnerabilidad a estimaciones extremas.

Por ejemplo, la figura 3 muestra un ejemplo de datos de entrenamiento proyectados en el plano PC3/PC4 del espacio de ascendencia (definido por los primeros cuatro PC de 1000 Genomas), donde los datos de entrenamiento carecen de individuos de ascendencia SAS. Los puntos y etiquetas rojos sólidos en la figura 3 muestran los centroides de cada superpoblación a partir de los datos de 1000 Genomas. Como puede verse, el entrenamiento carece de muestras alrededor del centroide de SAS.

La figura 4 muestra estimaciones del tamaño del efecto (el tamaño del efecto estimado por unidad de PRS) para un individuo en la ubicación promedio de cada una de las cinco superpoblaciones principales en 1000 Genomas bajo el modelo (3), donde el modelo se entrena usando los datos de entrenamiento de la figura 3. Las estimaciones del tamaño del efecto se notifican para cada superpoblación usando los primeros cuatro PC en una regresión logística (azul) y se comparan con la estimación discreta dentro de esa población en las pruebas (rojo). La estimación discreta es la estimación calculada usando únicamente los datos de prueba de esa superpoblación, en lugar de usar los datos de prueba completos de todas las superpoblaciones con las correcciones apropiadas para su posición en el espacio de

ascendencia. Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación de PC mediana de la superpoblación en 1000 Genomas, los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. La estimación del punto del cuadrado azul ("Promedio del grupo") representa la estimación del tamaño del efecto promedio entre los individuos del conjunto de pruebas dentro de cada etiqueta de superpoblación. Debido a la falta de individuos de SAS en los datos de entrenamiento, el tamaño del efecto estimado para un individuo en la posición SAS está por debajo del límite mínimo razonablemente esperado de 1,0.

La figura 5 muestra las estimaciones basadas en máxima verosimilitud (MLE) de los coeficientes de interacción ($\beta_{PRS \times PC_j}$) en el modelo (3)) y sus intervalos de confianza al 95 % asociados. Para cada coeficiente de interacción, la MLE se muestra como un punto negro y los errores estándar asociados como "bigotes". Como se esperaba, dada la escasez de datos de entrenamiento a lo largo de PC3 (debido a la falta de la superpoblación SAS), los errores estándar para el coeficiente de interacción en el ajuste logístico del modelo (3) son grandes. Esto conduce a la estimación del tamaño del efecto irrazonablemente baja para un individuo SAS en la figura 4, debido a la gran estimación puntual negativa de PC3 (para la cual los individuos del sur de Asia tienen la mayor ponderación) en la figura 5.

Para protegerse contra este tipo de estimación inestable, se pueden considerar una serie de alteraciones. Un enfoque es determinar el PC que separa la superpoblación que está ausente de los datos de entrenamiento de las otras superpoblaciones en el espacio de ascendencia y eliminar ese PC como covariable (y su correspondiente coeficiente de interacción $\beta_{PRS \times PC_j}$ con el PRS) en el modelo (3). Sin embargo, este enfoque está sesgado a sobreestimar el tamaño del efecto porque el modelo "piensa" que los nuevos individuos objetivo en la superpoblación faltante están mucho más cerca en el espacio de ascendencia de los individuos en los datos de entrenamiento de lo que realmente están.

Este enfoque de la corrección se muestra en la figura 6, continuando con los mismos datos usados en los ejemplos anteriores de las figuras 3-5. La figura 6 muestra las estimaciones del tamaño del efecto para un individuo en la ubicación promedio de cada superpoblación bajo el modelo (3) en un espacio de ascendencia definido por solo tres PC. Se notifican estimaciones del tamaño del efecto para cada una de las cinco superpoblaciones principales en 1000 Genomas usando los tres PC en una regresión logística de acuerdo con el modelo (3) después de eliminar el PC que distingue a los asiáticos del sur de las superpoblaciones restantes (azul) en un modelo con los primeros cuatro PC, y se compara con la estimación discreta dentro de esa población en las pruebas (rojo). Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación PC mediana de la superpoblación en 1000 Genomas y los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. La estimación del punto del cuadrado azul ("Promedio del grupo") representa la estimación del tamaño del efecto promedio entre individuos dentro de cada etiqueta de superpoblación. Como puede verse, los cocientes de probabilidades para SAS son más altos que la estimación discreta dentro de las superpoblaciones en las pruebas.

Un enfoque alternativo es introducir términos de regularización en el modelo de la dependencia de la contribución genética para protegerse contra el sobreajuste, por ejemplo, usando LASSO, red elástica o regresión de crestas.

Posición relativa

En alguna realización, calcular la contribución genética comprende calcular una distancia en el espacio de ascendencia entre la posición individual y una posición de referencia en el espacio de ascendencia, y calcular la contribución genética usando la distancia. En este caso, el tamaño del efecto se puede modelizar variando continuamente con la ascendencia usando el siguiente modelo:

$$Y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = \beta_0 + \beta_{PRS}X_{PRS,i} + \beta_{EUR}X_{EUR,i} + \beta_{PRS \times EUR}X_{PRS,i}X_{EUR,i} \quad (4)$$

dónde $X_{EUR,i}$ es la distancia genética desde el individuo i hasta la posición mediana de la superpoblación EUR en el espacio de ascendencia. En lo sucesivo, esta distancia se define como la distancia euclidiana siguiendo una escala de los primeros cuatro PC por la varianza que explican en los datos genéticos de referencia. En otras palabras, el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes; y calcular la contribución genética comprende ajustar a escala cada eje del espacio de ascendencia usando una varianza representada por el eje respectivo en los datos genéticos de referencia antes de calcular la distancia. Sin embargo, el método no se limita a estas opciones: se podrían incluir más o menos PC y se podría usar una medida de distancia no euclidiana.

La figura 7 muestra los mismos datos de entrenamiento de ejemplo usados en las figuras 3 a 6. En esta ocasión, los datos se proyectan en un plano en el espacio de ascendencia definido por PC1 y PC2. Cada punto es un individuo de entrenamiento en los datos de entrenamiento, coloreados por su tamaño del efecto de PRS estimado ("cociente de probabilidades") de acuerdo con el modelo (4). En los modelos lineales del presente método, no hay discontinuidades a diferencia del enfoque del estado de la técnica que se muestra en la figura 1.

El modelo (4) da como resultado una estrecha concordancia con las estimaciones del tamaño del efecto de SAS del conjunto de pruebas, como se muestra en la figura 8. En la figura 8, se notifican estimaciones del tamaño del efecto para cada una de las cinco superpoblaciones principales en 1000 Genomas usando los tres PC en una regresión logística de acuerdo con el modelo (4), y se comparan con la estimación discreta dentro de esa población en las pruebas (rojo). Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación PC mediana de la superpoblación en 1000 Genomas y los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. La estimación del punto del cuadrado azul ("Promedio del grupo") representa la estimación del tamaño del efecto promedio entre individuos dentro de cada etiqueta de superpoblación.

Sin embargo, las estimaciones de las pruebas podrían ser diferentes debido a efectos de cohorte/ambientales o una gran incertidumbre debido a tamaños de muestra pequeños (como en el caso de AMR_NAT). Cabe señalar que la estimación del tamaño del efecto de EAS es ligeramente menor y AFR_SS ligeramente mayor en comparación con los enfoques anteriores, lo que puede reflejar la naturaleza menos flexible de este modelo.

El modelo (4) (y otros modelos similares de "posición relativa") supone que el tamaño del efecto de PRS cambia de la misma manera cuando cambia la posición individual, independientemente del eje en el espacio de ascendencia a lo largo del cual se mueve la posición individual. Esta es una suposición razonable, porque se ha observado previamente que la discriminación/precisión de la predicción cae aproximadamente linealmente con la distancia en el espacio de ascendencia. Además, se observan tamaños de efecto similares al agrupar a los individuos por distancia.

El modelo (4) tiene las mismas ventajas que el modelo lineal multidimensional. Sin embargo, también tiene el beneficio adicional de que, dado que el espacio de ascendencia se reduce efectivamente a una única dimensión a los efectos del modelo de contribución genética, es mucho menos probable que un nuevo individuo ocupe un puesto en el régimen de extrapolación (es decir, fuera de la región cubierta por los datos de entrenamiento). Más bien, el modelo interpolará entre individuos de entrenamiento en los datos de entrenamiento individuales, lo que da como resultado estimaciones más estables de los parámetros del modelo (es decir, los coeficientes de interacción). Esto regulariza efectivamente el coeficiente de interacción.

Una limitación del modelo (4) es que hace una suposición adicional en comparación con el modelo multidimensional (ejemplificado por el modelo (3)). Esto significa que el modelo es menos flexible y tiene menos grados de libertad. Es posible que la suposición de que el tamaño del efecto de PRS cambia de la misma manera en todas las direcciones en el espacio de ascendencia sea inexacta. En este caso, el modelo de posición relativa estaría subajustando los datos de entrenamiento, dando como resultado estimaciones sesgadas. Además, el modelo de posición relativa también requiere la especificación de un punto de referencia. Existe un problema de extrapolación similar para este modelo si las muestras africanas no están presentes en los datos de entrenamiento usados para ajustar el modelo, aunque probablemente sea más robusto que el modelo multidimensional ante dicha ausencia de datos de entrenamiento.

Modelo jerárquico

Los modelos (3) y (4) pueden verse como extremos de un espectro. En un extremo del espectro (modelo (3)), cada coeficiente de interacción de PC se estima sin compartir información entre los ejes de variación en el espacio de ascendencia. En este caso, las subcontribuciones para cada eje del espacio de ascendencia son independientes entre sí. En el otro extremo del espectro (modelo (4)), los ejes del espacio de ascendencia se colapsan en una sola distancia. En efecto, esta distancia comparte completamente la información entre los PC.

Entre estos dos extremos hay un modelo "jerárquico", que comparte información entre los diferentes términos de interacción y PC, manteniendo la flexibilidad para permitir diferentes coeficientes para cada uno. Esto proporciona efectivamente un número de grados de libertad intermedios entre uno (como para los modelos de "distancia relativa") y el número de PC (para los modelos de "distancia absoluta"). En dichas realizaciones, al menos dos de las subcontribuciones tienen dependencias de las coordenadas respectivas de la posición individual que están relacionadas por una probabilidad a priori compartida. Estos modelos usan una probabilidad a priori compartida entre los coeficientes de interacción. Esto permite compartir información y representa la creencia previa de que el tamaño del efecto de PRS debería cambiar de manera similar a lo largo del eje en el espacio de ascendencia correspondiente a cada PC, permitiendo al mismo tiempo que el cambio pueda no ser idéntico a lo largo de cada eje. En algunas realizaciones, el modelo puede aprender de los datos de entrenamiento cuán similares son entre sí los cambios a lo largo de cada eje.

Relacionar las dependencias de las subcontribuciones de la probabilidad a priori compartida tiene cierta similitud con un modelo lineal jerárquico estándar (también conocido como modelo multinivel) que comparte información entre grupos de "individuos". Sin embargo, en el presente caso, la información se comparte a través de los propios coeficientes de interacción (que representan las dependencias de las subcontribuciones de las coordenadas respectivas) porque el modelo se ha movido al espacio de ascendencia continuo.

El presente modelo jerárquico también difiere de un modelo jerárquico estándar en que, dado que cada coeficiente de

interacción tiene el mismo número de puntos de datos, el modelo se basa en diferentes incertidumbres de los coeficientes de interacción para informar de en qué medida debe regularizarse cada coeficiente de interacción.

Para un PC que distingue una superpoblación que está presente en los datos genéticos de referencia usados para definir el espacio de ascendencia, pero que falta en los datos de entrenamiento, la estimación será más incierta debido al intervalo más pequeño sobre el cual se aprende el gradiente. Como resultado, estas estimaciones se regularizarán más fuertemente.

Este modelo se puede expresar de la siguiente manera:

$$\begin{aligned}
 Y_i &\sim \text{Bernoulli}(\pi_i) \\
 \text{logit}(\pi_i) &= \beta_0 + \beta_{PRS} X_{PRS,i} + \sum_{j=1}^k \left(\beta_{PC_j} X_{PC_j,i} + \beta_{PRS \times PC_j} X_{PRS,i} X_{PC_j,i} \right) \\
 \beta_{PRS \times PC_j} &= \sigma_{PRS \times PC} \beta_{PRS \times PC_j}^* + \mu_{PRS \times PC} \\
 \beta_{PRS \times PC_j}^* &\sim N(0, 1) \\
 \mu_{PRS \times PC} &\sim N(a, b) \\
 \sigma_{PRS \times PC} &\sim \text{semiCauchy}(c, d)
 \end{aligned} \tag{5}$$

lo que implica:

$$\beta_{PRS \times PC_j} \sim N(\mu_{PRS \times PC}, \sigma_{PRS \times PC}^2)$$

donde a , b , c y d son hiperparámetros.

En cuanto al modelo (3), cada subcontribución comprende un producto de la puntuación de riesgo poligénico y la coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. Sin embargo, ahora las subcontribuciones están relacionadas entre sí por la probabilidad a priori compartida.

En cuanto a los otros modelos, los datos y ejemplos proporcionados en las figuras solo usan los primeros cuatro componentes principales, pero en general se podría incluir cualquier número. Aumentar este número al número total de PC de señal en los datos genéticos de referencia podría ayudar en el aprendizaje del parámetro de varianza, $\sigma_{PRS \times PC}$.

Como los signos de los componentes principales son arbitrarios, estos deben armonizarse en relación con los datos de entrenamiento, de lo contrario el $\sigma_{PRS \times PC}$ aprendido se inflará artificialmente, por ejemplo, debido a que un coeficiente es -0,1 y el otro +0,1. Esto puede dar lugar a una pérdida de intercambio de información y a una falta de regularización suficiente. La armonización se puede realizar cambiando los signos de los componentes principales (por ejemplo, multiplicando por menos uno) de modo que los signos de todos los coeficientes de interacción aprendidos sean los mismos. Dado que los signos de los componentes principales son arbitrarios, no importa el signo elegido (positivo o negativo) siempre que todos sean coherentes.

El $\beta_{PRS \times PC_j}$ se muestrean a partir de una distribución, cuyos parámetros se aprenden usando información tanto de los datos en sí como de las probabilidades a priori sobre estos parámetros (que tienen hiperparámetros establecidos por el usuario). Es así como la probabilidad a priori compartida incide en la definición de las dependencias de las subcontribuciones. En efecto, la probabilidad a priori compartida se especifica de modo que las dependencias de las al menos dos subcontribuciones se muestrean a partir de la misma distribución. Se toma un conjunto de muestras de la distribución posterior para cada componente principal, no para cada individuo objetivo o individuo de entrenamiento. El resultado es una distribución posterior para cada $\beta_{PRS \times PC_j}$.

La probabilidad a priori compartida se determina usando datos de entrenamiento de una pluralidad de individuos de entrenamiento y uno o más hiperparámetros predeterminados. La fuerza de la creencia previa de que los coeficientes de interacción deberían ser similares se puede variar variando los hiperparámetros. En la implementación específica ejemplificada anteriormente, el hiperparámetro d se puede variar, con valores más pequeños de d correspondientes a una creencia más fuerte de que son similares.

A partir de la distribución posterior de los coeficientes de interacción, $f\beta_{PRS \times PC_j}$, se puede calcular una distribución del cociente de probabilidades para cada individuo de entrenamiento (o el individuo objetivo). Cada muestra posterior para un individuo i es la suma del producto de los predictores (con $X_{PRS,i} = 1$ para obtener el tamaño del efecto por unidad de PRS) y las estimaciones posteriores de los coeficientes, es decir,

$$\beta_{PRS} X_{PRS,i} + \sum_{j=1}^k$$

($\beta_{PCj} X_{PCj,i} + \beta_{PRS \times PCj} X_{PRS,i} X_{PCj,i}$). De modo que las distribuciones completas están dadas por $A \cdot B$ donde A es una matriz $N \times M$ para N individuos y M predictores [$X_{PRS,i}$, $X_{PRS,i} X_{PC1,i}$, $X_{PRS,i} X_{PC2,i}$, ...] (como anteriormente, con $X_{PRS,i}$ establecido a 1 para obtener una relación de probabilidades "por 1 unidad de PRS") y B es una matriz $M \times P$, donde P es el número de muestras posteriores.

A continuación se puede derivar una estimación puntual para cada individuo, dada alguna función de pérdida, error cuadrático medio comúnmente, pero podría ser otro, por ejemplo, pérdida exponencial asimétrica si la sobreestimación se considera peor que la subestimación o viceversa.

El modelo se muestra en una versión "descentrada" que ayuda a ajustar los parámetros del modelo, especialmente el gran $\sigma_{PRS \times PC}$. Sin embargo, el modelo podría formularse en una versión "centrada" que sea equivalente y se muestre como un modelo implícito. Los coeficientes de PC de no interacción, β_{PCj} no afecta directamente al tamaño del efecto, sino más bien la probabilidad inicial absoluta en un área del espacio de ascendencia. Esto significa que el argumento a favor de que son similares es más débil que el de los coeficientes de interacción $\beta_{PRS \times PCj}$. Sin embargo, los coeficientes de no interacción también podrían estar relacionados mediante una probabilidad a priori jerárquica. Cuando no se indique ningún precedente en los modelos anteriores, se supone una probabilidad a priori uniforme o una probabilidad a priori débilmente informativa.

Los parámetros de este modelo bayesiano (5) se ajustan usando el Monte Carlo hamiltoniano para obtener los resultados de ejemplo que se muestran en las figuras. Sin embargo, también se podrían usar otras aproximaciones, por ejemplo, inferencia variacional o aproximación anidada integrada de Laplace. Cuando se instala usando Monte Carlo hamiltoniano, el modelo es más lento que los otros enfoques lineales, aunque más rápido que el proceso gaussiano variacional que se analiza adicionalmente más adelante.

La figura 9 muestra estimaciones de tamaño del efecto para un individuo en la ubicación promedio de cada superpoblación bajo el modelo (5). Las estimaciones de tamaño del efecto se notifican para cada una de las cinco superpoblaciones principales en 1000 Genomas usando el modelo (5) (azul) y se comparan con la estimación discreta dentro de esa población en las pruebas (rojo). Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación PC mediana de la superpoblación en 1000 Genomas y los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. La estimación del punto del cuadrado azul ("Promedio del grupo") representa la estimación del tamaño del efecto promedio entre individuos dentro de cada etiqueta de superpoblación.

El modelo (5) regulariza lo suficiente como para generar un tamaño del efecto de SAS muy cercano a la estimación discreta dentro del conjunto de pruebas, como en el modelo unidimensional anterior. Sin embargo, el modelo (5) también es lo suficientemente flexible como para tener una estimación de tamaño del efecto para EAS/AFR_SS similar a la estimación discreta dentro del conjunto de pruebas, como en los otros modelos multidimensionales. El modelo (5) logra un equilibrio entre los modelos unidimensionales (posición relativa) y los modelos multidimensionales (posición absoluta) al aprender cuán similares son los coeficientes de interacción a partir de los datos de entrenamiento, en lugar de codificar un efecto idéntico o permitir que los efectos sean "libres". Si los datos de entrenamiento sugieren que las dependencias de las subcontribuciones a lo largo de diferentes ejes del espacio de ascendencia son realmente diferentes, entonces se aprenderá una desviación estándar grande y las dependencias no se regularizarán con tanta fuerza.

Modelos no lineales

Todos los modelos analizados hasta ahora han sido modelos lineales, donde las dependencias de la contribución genética (o sus subcontribuciones) de la posición individual en el espacio de ascendencia han sido lineales. Esto significa que un cambio de cierto tamaño y dirección en la posición individual tiene el mismo efecto sobre la contribución genética independientemente de en qué parte del espacio de ascendencia ocurra. Sin embargo, también se pueden considerar modelos no lineales. En dichas realizaciones, calcular la contribución genética comprende usar una dependencia no lineal de la posición individual.

Modelos aditivos generalizados (GAM)

Un primer ejemplo de modelo no lineal es un modelo aditivo generalizado (GAM). En el caso de un GAM, en lugar de tener una dependencia lineal simple entre cada coeficiente y covariable (PRS o posición individual), se puede modelizar esta dependencia como cualquier función de cada covariable:

$$Y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(p_i|X) = \beta_0 + f_1(X_1) + f_2(X_2) + \dots + f_M(X_M) \quad (6)$$

La dependencia no lineal puede comprender una función regularizada. Esto asegura que la dependencia sea fluida y no contenga discontinuidades repentinas u otras características que probablemente no sean biológicamente razonables. Una forma funcional común para elegir son las B splines penalizadas (también conocidas como P-splines), que permiten modelizar formas no lineales. En dicha realización, la dependencia no lineal comprende una B-spline (línea polinómica suave básica) penalizada. En una B-spline penalizada, se aplica una penalización ("lambda") a la curvatura de la función, con la función volviéndose completamente lineal con valores mayores. Este valor lambda se puede aprender usando validación cruzada sobre una cuadrícula de valores. Una forma de elegir el valor lambda preferido es elegir el valor que minimice el estimador de riesgo insesgado.

Como ejemplo de este tipo de modelo, la figura 10 muestra estimaciones de tamaño del efecto para un individuo en la ubicación promedio de cada superpoblación bajo el modelo (6). Un modelo de posición relativa unidimensional se usa con una única distancia desde EUR y su interacción con PRS. Las estimaciones de tamaño del efecto se notifican para cada una de las cinco superpoblaciones principales en 1000 Genomas usando el modelo (6) (azul) y se comparan con la estimación discreta dentro de esa población en las pruebas (rojo). Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación de PC mediana de la superpoblación en 1.000 Genomas, los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. Los puntos cuadrados azules ("Promedio del grupo") indican la estimación del tamaño del efecto promedio entre los individuos del conjunto de pruebas con cada etiqueta de superpoblación.

Los resultados que usan el modelo (6) son muy similares a los resultados del modelo unidimensional que se muestran en la figura 8. El modelo (6) es más rápido que la estrategia del proceso gaussiano variacional que se analiza a continuación, pero aún permite de manera flexible dependencias no lineales. Sin embargo, la búsqueda de cuadrícula en diferentes valores lambda para ajustarse al ejemplo de B-spline penalizada en la figura 10 es lenta, lo que limita el espaciado de la cuadrícula que se puede usar para probar diferentes valores lambda.

Procesos gaussianos

Una forma alternativa de ajustar una dependencia no lineal es usar procesos gaussianos. En dichas realizaciones, la dependencia no lineal se determina usando un proceso gaussiano como distribución previa para calcular la contribución genética mediante inferencia bayesiana.

En el caso de regresión de proceso gaussiano, se supone la siguiente configuración. Los datos de entrenamiento comprenden una colección de observaciones $[y_1, y_2, \dots, y_n]$ de alguna función desconocida $f: X \rightarrow R$ en una colección de ubicaciones $[X_1, X_2, \dots, X_n]$ con ruido añadido:

$$y_i = f(x_i) + \varepsilon_i \text{ for } i = 1, 2, \dots, n.$$

dónde ε_i es un término de ruido.

A partir de estos datos de entrenamiento, una estimación de la función desconocida f para los nuevos datos que llegan: $y^* = f(x^*) + \varepsilon^*$ está por determinar. Para ello, se supone que los datos se muestrean a partir de un proceso estocástico, conocido como proceso gaussiano (GP). Las muestras tomadas de un GP tienen funciones $\mu(\cdot)$, y $k(\cdot, \cdot)$ de modo que la función de expectativa $E[f(x)] = \mu(x)$ y la función de covarianza $Cov[f(x), f(x')] = k(x, x')$. La abreviatura de este procedimiento de muestreo es $GP(\mu(\cdot), k(\cdot, \cdot))$.

Cada subconjunto finito de un GP (en particular el subconjunto de datos que se usa para entrenamiento) se muestrea a partir de una distribución gaussiana multivariada. La forma de interpolar entre los puntos de los datos de entrenamiento está determinada por la elección de las funciones. $\mu(\cdot)$, y $k(\cdot, \cdot)$. En la presente situación, existe la complejidad añadida de que Y no es continua, sino una etiqueta de caso/control discreta. Se supone que Y tiene distribución de Bernoulli, y la distribución latente subyacente que rige el riesgo de ser un caso se muestrea a partir de un GP.

$$Y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = GP(\mu(\cdot), k(\cdot, \cdot)) \quad (7)$$

También existe el problema adicional de que los métodos estándar para evaluar las medias y varianzas predictivas tienen complejidad computacional. $O(n^3)$, donde n es el número de puntos de datos. Esto no es un problema cuando n es del orden de miles, pero hace que los enfoques estándar sean intratables cuando los tamaños de las muestras se acercan a cientos de miles, como puede ser el caso de los datos genéticos.

Para estimar el tamaño del efecto por PRS para un individuo, se usa la distribución posterior. Se determina una distribución posterior para la inferencia bayesiana usando el proceso gaussiano y datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes.

Un problema con la introducción de una función de enlace (la función de Bernoulli) que relaciona el espacio latente

con el estado de casos y controles Y es que la distribución posterior ya no es gaussiana. Para mantener la manejabilidad, en algunas realizaciones, la posterior no gaussiana se aproxima con una gaussiana. Por lo tanto, determinar la distribución posterior comprende aproximar la distribución posterior como una distribución gaussiana. Existen una serie de métodos que se pueden usar para realizar esta aproximación. Por ejemplo, una aproximación de Laplace, propagación de expectativas y minimización de la divergencia Kullback-Leibler (KL). En la presente realización, se usa este último enfoque, aproximando la distribución posterior exacta p por una gaussiana q , y minimizando $KL(q||p)$. Este problema de minimización se puede resolver usando el método de Newton y también es $O(n^3)$. Los detalles completos de la aproximación variacional se describen en detalle en (Nickisch y Rasmussen 2008).

Cuando se usan procesos gaussianos, es común asumir una función media nula $\mu(.) = 0$, de modo que el proceso gaussiano viene dado por $GP(0, k(.,.))$. Esta elección simplifica el ajuste y la media siempre se puede volver a añadir después, ya que la forma del proceso gaussiano está completamente determinada por su función de covarianza. En dicha realización, el proceso gaussiano tiene una función media nula. Sin embargo, en algunos casos, puede ser deseable incluir una función media distinta de cero, ya sea antes del ajuste o añadido después del ajuste. Por ejemplo, se puede usar una función media distinta de cero para controlar el OR (tamaño del efecto) por unidad de PRS en regiones del espacio de ascendencia alejadas de los datos de entrenamiento, efectivamente como una probabilidad a priori sobre la contribución de PRS en ausencia de datos genéticos en esa ubicación ancestral. Por ejemplo, el tamaño del efecto observado en individuos de ascendencia africana podría establecerse como el tamaño del efecto mínimo en todo el espacio de ascendencia, sobre la base de que representan el "peor de los casos" para la atenuación de la ascendencia. Por lo tanto, en algunas realizaciones, el proceso gaussiano tiene un vector medio correspondiente a una estimación previa de la contribución genética al riesgo para el individuo objetivo. Este enfoque puede hacer que la adaptación sea más desafiante debido a la no linealidad del GP.

Lo más importante para el comportamiento del proceso gaussiano es la elección de la función núcleo $k(x, x')$. La función núcleo describe una forma funcional para la covarianza esperada entre cualquier par de ubicaciones x y x' .

Opcionalmente, la función núcleo del proceso gaussiano es una función estacionaria. Una función estacionaria depende de la distancia entre dos puntos y no de sus posiciones absolutas. Esto significa que el comportamiento del proceso gaussiano es similar en todo el espacio de ascendencia. La función núcleo también puede ser isotrópica, donde depende sólo de la magnitud de la distancia entre dos puntos. Opcionalmente, la función núcleo del proceso gaussiano decae a cero al disminuir la similitud entre muestras. La elección de una función núcleo en decadencia garantiza que el valor de la función vuelva a decaer a la media en regiones del espacio de ascendencia alejadas de cualquier dato de entrenamiento. Esto puede ayudar a controlar y reducir el comportamiento inesperado de la función cuando un individuo objetivo tiene una posición individual que no está muy cerca de los datos de entrenamiento en el espacio de ascendencia. Por ejemplo, la función núcleo puede ser una función de base radial o una función de covarianza cuadrática racional. En las realizaciones de ejemplo siguientes, se usa una función de base radial (RBF) como núcleo:

Opcionalmente, la función núcleo del proceso gaussiano depende de uno o más hiperparámetros. Por ejemplo, los hiperparámetros pueden comprender un hiperparámetro asociado con cada una de la puntuación de riesgo poligénico, la posición individual, y una interacción entre la puntuación de riesgo poligénico y la posición. En los ejemplos a continuación, se usan un hiperparámetro de varianza θ y tres hiperparámetros de escala de longitud: l_{PRS} , l_{PC} , $l_{PRS \times PC}$ para las interacciones de la PRS, los PC y PRS $\times$ PC, respectivamente.

Los hiperparámetros pueden comprender hiperparámetros asociados con una coordenada de la posición individual a lo largo de cada eje del espacio de ascendencia, es decir, correspondiente a cada uno de los PC. Los hiperparámetros pueden comprender hiperparámetros asociados con una interacción entre la PRS y cada coordenada de la posición individual. Sin embargo, en este ejemplo, se usa un único hiperparámetro de escala de longitud para la posición individual (los PC) y un único hiperparámetro de escala de longitud para la interacción entre la PRS y la posición individual (PRS $\times$ PC). Esto se logra ajustando a escala las escalas de longitud relativa de los PC según la varianza que explican en los datos genéticos de referencia. Por lo tanto, el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes (como se analizó anteriormente), y calcular la contribución genética comprende ajustar a escala cada eje del espacio de ascendencia usando una varianza representada por el eje respectivo en los datos genéticos de referencia antes de calcular la distancia. Este ajuste a escala y uso de un número reducido de hiperparámetros aumenta la velocidad, protege contra el sobreajuste y refuerza una noción sólida de "distancia" en el espacio de ascendencia.

En cuanto a los otros modelos analizados anteriormente, los procesos gaussianos se pueden usar con un enfoque multidimensional de posición absoluta, o un enfoque unidimensional de posición relativa.

Posición absoluta

En los modelos de posición absoluta, la contribución genética comprende una suma de subcontribuciones para cada eje del espacio de ascendencia, cada subcontribución calculada usando una coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia. Para aplicar el marco del proceso gaussiano al problema de predicción en este caso, la muestra correspondiente a cada uno de los individuos en formación está definida por un

vector:

$$x_i = [x_{i,PRS}, x_{i,PC}, x_{i,PRS \times PC}]$$

- 5 dónde $x_{i,PRS}$, $x_{i,PC}$ y $x_{i,PRS \times PC}$ son la PRS estandarizada, el vector que define la posición en la ubicación del espacio de ascendencia y la interacción correspondiente para el individuo i , respectivamente. La función núcleo que usa una función de base radial viene dada por:

$$k(x_i, x_j) = \theta \exp \left(-\frac{1}{2} \left[\frac{|x_{i,PRS} - x_{j,PRS}|^2}{l_{PRS}^2} + \frac{|x_{i,PC} - x_{j,PC}|^2}{l_{PC}^2} + \frac{|x_{i,PRS \times PC} - x_{j,PRS \times PC}|^2}{l_{PRS \times PC}^2} \right] \right).$$

- 10 Apelar a la categorización de procesos gaussianos permite incorporar naturalmente los efectos no lineales de los PC en la predicción del estado de los casos individuales. Asimismo, dependiendo de la forma funcional del núcleo, el proceso gaussiano puede imponer que el tamaño del efecto de la PRS sea 0 lejos de los datos de entrenamiento. Esta es una propiedad deseable: en ausencia de información (según lo estipulado por el núcleo), se espera que la estimación sea la prevalencia de la enfermedad en los datos de entrenamiento.

- 15 El beneficio de emplear la categorización de procesos gaussianos también es un inconveniente. En este marco, los cambios en PRS pueden afectar de forma no lineal a la probabilidad de los casos. Por tanto, el impacto de la PRS sobre el riesgo varía en función de su valor (a menos que se aplique un núcleo lineal para los términos que incluyen la PRS). La búsqueda de un GP también puede ser lenta. La velocidad se puede mejorar usando procesos gaussianos variacionales escasos o realizando una aproximación mediante una reducción de dimensionalidad. El primer enfoque puede resultar difícil de optimizar de forma robusta. Sin embargo, artículos recientes han sugerido enfoques para aumentar la robustez en la regresión del proceso gaussiano, pero aún no para la categorización. La función núcleo de base radial es una opción apropiada, porque los hiperparámetros de la escala de longitud entrenados dan como resultado estimaciones que rápidamente regresan a un nivel inicial en ausencia de datos.

Posición relativa

- 20 En el enfoque de posición relativa, calcular la contribución genética comprende calcular una distancia en el espacio de ascendencia entre la posición individual y una posición de referencia en el espacio de ascendencia, y calcular la contribución genética usando la distancia. Para esta implementación unidimensional del marco de GP, se utiliza un conjunto diferente de predictores. En lugar de los predictores del modelo (4), es decir,

$$Y_i \sim \text{Bernoulli}(\pi_i) \\ \text{logit}(\pi_i) = \beta_0 + \beta_{PRS} X_{PRS,i} + \beta_{EUR} X_{EUR,i} + \beta_{PRS \times EUR} X_{PRS,i} X_{EUR,i} \quad (4)$$

se usa un proceso gaussiano de acuerdo con el modelo (7) que tiene una función media nula y la siguiente función núcleo:

$$k(x_i, x_j) = \theta \exp \left(-\frac{1}{2} \left[\frac{|x_{i,PRS} - x_{j,PRS}|^2}{l_{PRS}^2} + \frac{|x_{i,EUR} - x_{j,EUR}|^2}{l_{EUR}^2} + \frac{|x_{i,PRS \times EUR} - x_{j,PRS \times EUR}|^2}{l_{PRS \times EUR}^2} \right] \right). \quad (8)$$

- 40 Como en la realización multidimensional anterior, se usan cuatro hiperparámetros en la función núcleo. En esta realización, estos se denotan θ , l_{PRS} , l_{EUR} y $l_{PRS \times EUR}$. Estos hiperparámetros rigen la escala general y las escalas de longitud sobre las cuales la información sobre la PRS, la distancia desde el punto de referencia (en este ejemplo, la distancia euclidiana desde el centro del agrupamiento europeo determinada como la posición del europeo medio en los datos genéticos de referencia), y la interacción entre la PRS y la distancia, respectivamente, se comparten entre pares de puntos.

- 50 La figura 11 muestra estimaciones del tamaño del efecto para un individuo en la ubicación promedio de cada superpoblación. Las estimaciones del tamaño del efecto se derivan de un marco de GP, usando el modelo (7), una función media nula y el núcleo en (8). Se notifican estimaciones del tamaño del efecto para cada una de las cinco superpoblaciones principales en 1000 Genomas usando una única distancia desde el centro del grupo europeo (como se ha definido anteriormente) y su interacción con PRS en un marco de GP. Las estimaciones ajustadas en los centroides de las superpoblaciones (azul) se comparan con la estimación discreta dentro de esa población en las pruebas (rojo). Los círculos azules ("1KG") se refieren al cociente de probabilidades en la ubicación PC mediana de la superpoblación en 1000 Genomas y los triángulos azules ("Empírico") se refieren al cociente de probabilidades en la

ubicación media de PC de los individuos del conjunto de pruebas dentro de una superpoblación determinada. La estimación del punto del cuadrado azul ("Promedio del grupo") representa la estimación del tamaño del efecto promedio entre individuos dentro de cada etiqueta de superpoblación.

- 5 Esta implementación de GP del modelo de posición relativa unidimensional permite cambios no lineales en los tamaños del efecto de PRS en función de la distancia desde el punto de referencia, es decir, incorpora la relación observada empíricamente entre la heredabilidad explicada por PRS y la distancia genómica desde el punto de referencia, pero no obliga a que esta relación sea lineal. Al igual que con los otros modelos espaciales de ascendencia unidimensionales (posición relativa) considerados en el presente documento, el modelo supone que el cambio en la heredabilidad explicado por la PRS es independiente de la dirección en la que nos movemos desde el punto de referencia en el espacio de ascendencia. Una desventaja de esto es que esta suposición puede tener un impacto indebido sobre las estimaciones del tamaño del efecto para las regiones del espacio de ascendencia para las cuales hay datos de entrenamiento disponibles, y puede proporcionar estimaciones del tamaño del efecto demasiado fiables en regiones donde hay pocos datos de entrenamiento disponibles.

15 Contribuciones que no dependen tanto de la PRS ni de la posición

- Todos los modelos de ejemplo presentados en el presente documento son susceptibles de extensión. Por ejemplo, se pueden incorporar más PRS entrenadas en ascendencias distintas o similares añadiendo nuevas PRS estandarizadas y términos de interacción a los modelos. A continuación se puede evaluar la contribución genética agregada de estas PRS al riesgo para el individuo objetivo.

- Los componentes de la contribución genética (o sus subcontribuciones) que dependen tanto de la PRS como de la posición individual se denominan generalmente términos de interacción en lo anterior. Además, como se muestra en los diversos modelos anteriores, la contribución genética puede incluir componentes que no dependen tanto de la PRS como de la posición individual. La contribución genética puede comprender un componente independiente de la ascendencia calculado usando la puntuación de riesgo poligénico que no depende de la posición individual. Por ejemplo, el componente independiente de la ascendencia se denota $\beta_{PRS}X_{PRS,i}$ en el modelo (4) anterior. De manera análoga, la contribución genética puede comprender un componente dependiente de la ascendencia calculado basándose en la posición individual que no depende de la puntuación de riesgo poligénico. Por ejemplo, el componente dependiente de la ascendencia se denota $\beta_{EUR}X_{EUR,i}$ en el modelo (4) anterior.

Otras características

- Como se ha mencionado anteriormente, se determina una dependencia de la contribución genética de la posición individual y la puntuación de riesgo poligénico usando datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes. Los datos de entrenamiento comprenden, para cada uno de los individuos de entrenamiento, datos genéticos, y si el individuo de entrenamiento tiene el fenotipo o la combinación de fenotipos objetivo. Los datos de entrenamiento pueden diferir de los datos genéticos de referencia usados para definir el espacio de ascendencia.

Herramienta de riesgo integrada

- En algunas realizaciones, los datos de entrenamiento comprenden además, para cada uno de los individuos de entrenamiento, datos informativos sobre una o más covariables no genéticas, y la contribución genética se estima conjuntamente en presencia de las covariables no genéticas. Por ejemplo, las covariables no genéticas pueden incluir una o más de peso, altura, características de comportamiento, rasgos médicos y otros biomarcadores, tales como mediciones basadas en sangre u orina.
- Esto permite que el modelo tenga en cuenta las correlaciones entre covariables genéticas y no genéticas al determinar la dependencia de la contribución genética, lo que puede mejorar aún más la precisión. En dicha realización, el método comprende además recibir datos de covariables individuales para el individuo objetivo, los datos de covariables individuales informativos sobre las covariables no genéticas adicionales para el individuo objetivo. Calcular la contribución genética puede comprender además el uso de los datos de covariables individuales. Como alternativa o adicionalmente, los datos de covariables individuales pueden usarse para calcular una contribución no genética al riesgo para el individuo objetivo.

Resultados

- Una vez calculada, se puede generar la contribución genética al riesgo. En algunas realizaciones, el método comprende además generar la contribución genética al riesgo. Como alternativa o adicionalmente, la contribución genética se puede usar como parte de cálculos adicionales que se usan para calcular otras medidas útiles para evaluar el riesgo de un individuo de desarrollar un fenotipo o combinación de fenotipos particular.
- En algunas realizaciones, el riesgo es un riesgo relativo en relación con un individuo que tiene una contribución genética estimada promedio. Esta contribución genética estimada promedio puede basarse en datos genéticos

promedio de individuos similares, por ejemplo, individuos que tienen ascendencia similar. La contribución genética estimada promedio también puede basarse en una prevalencia promedio del fenotipo o combinación de fenotipos en la población o en individuos similares. En dichas realizaciones, el método comprende además calcular S50 el riesgo relativo para el individuo objetivo para el fenotipo objetivo o combinación de fenotipos objetivo usando la contribución genética calculada usando cualquiera de los métodos descritos anteriormente de cualquier reivindicación anterior y una contribución no genética al riesgo relativo, y generar S60 el riesgo relativo. Como se ha mencionado anteriormente, la contribución no genética puede calcularse basándose en covariables no genéticas.

- 5
- 10

Calcular el riesgo relativo puede comprender usar una función de pérdida para determinar un valor del riesgo relativo para el individuo objetivo a partir de una distribución del riesgo relativo para el individuo objetivo. Por ejemplo, la función de pérdida puede ser una función de error cuadrático medio o una función de pérdida exponencial asimétrica.

Además del riesgo relativo, el método se puede usar para calcular el riesgo absoluto. En algunas realizaciones, el método comprende además calcular un riesgo absoluto para el individuo objetivo para el fenotipo objetivo o combinación de fenotipos objetivo usando la contribución genética calculada usando el método de cualquier reivindicación anterior, y generar el riesgo absoluto.

- 15

REIVINDICACIONES

1. Un método implementado por ordenador para analizar datos genéticos, que comprende:

- 5 recibir una puntuación de riesgo poligénico (S10) para un fenotipo objetivo o una combinación de fenotipos objetivo para un individuo objetivo;
recibir datos genéticos individuales (S20) para el individuo objetivo, los datos genéticos individuales (20) informativos sobre una ascendencia del individuo objetivo;
determinar una posición individual (S30) en un espacio de ascendencia usando los datos genéticos individuales (20), la posición individual representada por una combinación de variables continuas o pseudocontinuas, o comprendiendo la posición individual una asignación a una combinación ponderada de una pluralidad de ascendencias; y
10 calcular una contribución genética (S40) a un riesgo para el individuo objetivo para el fenotipo objetivo o combinación de fenotipos objetivo usando la puntuación de riesgo poligénico (10) y la posición individual, teniendo la contribución genética al riesgo una dependencia continua o pseudocontinua de la posición individual, en donde el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes.

2. El método de la reivindicación 1, en donde la contribución genética comprende una suma de subcontribuciones para cada eje del espacio de ascendencia, cada subcontribución calculada usando una coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia, opcionalmente en donde

- 25 i) el espacio de ascendencia es no isotrópico, de modo que la dependencia de cada subcontribución de la coordenada respectiva de la posición individual difiere entre las subcontribuciones, opcionalmente en donde al menos dos de las subcontribuciones tienen dependencias de las coordenadas respectivas de la posición individual que están relacionadas por una probabilidad a priori compartida, opcionalmente además en donde la probabilidad a priori compartida se especifica de modo que las dependencias de las al menos dos subcontribuciones se muestrean a partir de la misma distribución, opcionalmente además en donde la probabilidad a priori compartida se determina usando datos de entrenamiento de una pluralidad de individuos de entrenamiento y uno o más hiperparámetros predeterminados; y/o
30 ii) cada subcontribución comprende un producto de la puntuación de riesgo poligénico y la coordenada de la posición individual a lo largo del eje respectivo del espacio de ascendencia.

3. El método de la reivindicación 1 o 2, en donde calcular la contribución genética comprende calcular una distancia en el espacio de ascendencia entre la posición individual y una posición de referencia en el espacio de ascendencia, y calcular la contribución genética usando la distancia, opcionalmente en donde:

- 40 i) la posición de referencia es una posición en el espacio de ascendencia de una ascendencia usada para entrenar coeficientes usados para calcular la puntuación de riesgo poligénico; y/o
ii) el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes; y
calcular la contribución genética comprende ajustar a escala cada eje del espacio de ascendencia usando una varianza representada por el eje respectivo en los datos genéticos de referencia antes de calcular la distancia; y/o
45 iii) la distancia es una distancia euclidiana en el espacio de ascendencia; y/o
iv) la contribución genética comprende un producto de la puntuación de riesgo poligénico y la distancia.

4. El método de cualquier reivindicación anterior, en donde calcular la contribución genética comprende:

- 50 i) usar una dependencia lineal de la posición individual; o
ii) usar una dependencia no lineal de la posición individual, opcionalmente en donde la dependencia no lineal comprende una función regularizada y/o una B-spline penalizada.

5. El método de cualquiera de las reivindicaciones 1 a 3, en donde calcular la contribución genética comprende usar una dependencia no lineal de la posición individual, y la dependencia no lineal se determina usando un proceso gaussiano como distribución previa para calcular la contribución genética usando inferencia bayesiana, opcionalmente en donde:

- 60 i) el proceso gaussiano tiene una función media nula, o el proceso gaussiano tiene un vector medio correspondiente a una estimación previa de la contribución genética al riesgo para el individuo objetivo; y/o
ii) una función núcleo del proceso gaussiano es una función estacionaria; y/o
iii) una función núcleo del proceso gaussiano decae a cero al disminuir la similitud entre muestras; y/o
iv) la función núcleo es una función de base radial o una función de covarianza cuadrática racional; y/o
v) se determina una distribución posterior para la inferencia bayesiana usando el proceso gaussiano y datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes, opcionalmente, en donde determinar la distribución posterior comprende aproximar la distribución

posterior como una distribución gaussiana.

6. El método de la reivindicación 5, en donde una función núcleo del proceso gaussiano depende de uno o más hiperparámetros, opcionalmente en donde los hiperparámetros comprenden un hiperparámetro asociado con cada una de la puntuación de riesgo poligénico, la posición individual, y una interacción entre la puntuación de riesgo poligénico y la posición.

7. El método de cualquier reivindicación anterior, en donde:

- i) la contribución genética comprende un componente independiente de la ascendencia calculado usando la puntuación de riesgo poligénico que no depende de la posición individual; y/o
- ii) la contribución genética comprende un componente dependiente de la ascendencia calculado basándose en la posición individual que no depende de la puntuación de riesgo poligénico; y/o
- iii) cada individuo de referencia se asigna a una de una pluralidad de ascendencias.

8. El método de cualquier reivindicación anterior, en donde un sistema de coordenadas del espacio de ascendencia se determina aplicando reducción de dimensiones a los datos genéticos de referencia, opcionalmente en donde:

- i) la reducción de dimensiones comprende análisis de componentes principales, análisis de componentes independientes, factorización matricial no negativa, o análisis factorial, o la reducción de dimensiones comprende una discretización del espacio de ascendencia en un conjunto finito de ascendencias, y la posición individual comprende una proporción de pertenencia continua o pseudocontinua para cada ascendencia del conjunto finito de ascendencias; y/o
- ii) el sistema de coordenadas del espacio de ascendencia se elige para maximizar la varianza en los datos genéticos de referencia representados por el espacio de ascendencia.

9. El método de cualquier reivindicación anterior, en donde el espacio de ascendencia tiene una dimensionalidad menor que la de los datos genéticos individuales, y determinar la posición individual comprende proyectar los datos genéticos individuales en el espacio de ascendencia.

10. El método de cualquier reivindicación anterior, en donde se determina una dependencia de la contribución genética de la posición individual y la puntuación de riesgo poligénico usando datos de entrenamiento de una pluralidad de individuos de entrenamiento que tienen una pluralidad de ascendencias diferentes, comprendiendo los datos de entrenamiento, para cada uno de los individuos de entrenamiento, datos genéticos, y si el individuo de entrenamiento tiene el fenotipo o la combinación de fenotipos objetivo,

opcionalmente en donde: los datos de entrenamiento comprenden además, para cada uno de los individuos de entrenamiento, datos informativos sobre una o más covariables no genéticas, y la contribución genética se estima conjuntamente en presencia de las covariables no genéticas; y el método comprende además recibir datos de covariables individuales para el individuo objetivo, los datos de covariables individuales informativos sobre las covariables no genéticas adicionales para el individuo objetivo, además opcionalmente en donde las covariables no genéticas incluyen una o más de peso, altura, características de comportamiento, rasgos médicos y otros biomarcadores, tales como mediciones basadas en sangre u orina.

11. El método de cualquier reivindicación anterior, en donde el riesgo es un riesgo relativo en relación con un individuo que tiene una contribución genética estimada promedio, y el método comprende además:

calcular el riesgo relativo para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la contribución genética y una contribución no genética al riesgo relativo; y
generar el riesgo relativo y/o la contribución genética al riesgo,
opcionalmente, en donde calcular el riesgo relativo comprende usar una función de pérdida para determinar un valor del riesgo relativo para el individuo objetivo a partir de una distribución del riesgo relativo para el individuo objetivo,
además, opcionalmente, la función de pérdida es una función de error cuadrático medio o una función de pérdida exponencial asimétrica.

12. El método de cualquier reivindicación anterior, que comprende, además:

calcular un riesgo absoluto para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la contribución genética; y
generar el riesgo absoluto y/o la contribución genética al riesgo.

13. El método de la reivindicación 11 o 12, en donde calcular el riesgo absoluto o relativo comprende determinar un cociente de riesgos instantáneos para el individuo objetivo, normalizándose el cociente de riesgos instantáneos usando la puntuación de riesgo poligénico y la contribución genética.

14. Un programa informático o un medio legible por ordenador que comprende instrucciones que, cuando el programa es ejecutado por un ordenador, hace que el ordenador lleve a cabo el método de cualquiera de las reivindicaciones 1 a 13.

5 15. Un aparato para analizar datos genéticos que comprende un procesador configurado para:

recibir una puntuación de riesgo poligénico para un fenotipo objetivo o una combinación de fenotipos objetivo para un individuo objetivo;

10 recibir datos genéticos individuales para el individuo objetivo, los datos genéticos informativos sobre una ascendencia del individuo objetivo;

determinar una posición individual en un espacio de ascendencia usando los datos genéticos individuales; y

calcular una contribución genética a un riesgo para el individuo objetivo para el fenotipo objetivo o la combinación de fenotipos objetivo usando la puntuación de riesgo poligénico y la posición individual, en donde:

15 la posición individual está representada por una combinación de variables continuas o pseudocontinuas, o la posición individual comprende una asignación a una combinación ponderada de una pluralidad de ascendencias;

el espacio de ascendencia se define usando datos genéticos de referencia de una pluralidad de individuos de referencia que tienen una pluralidad de ascendencias diferentes; y

20 la contribución genética tiene una dependencia continua o pseudocontinua de la posición individual.

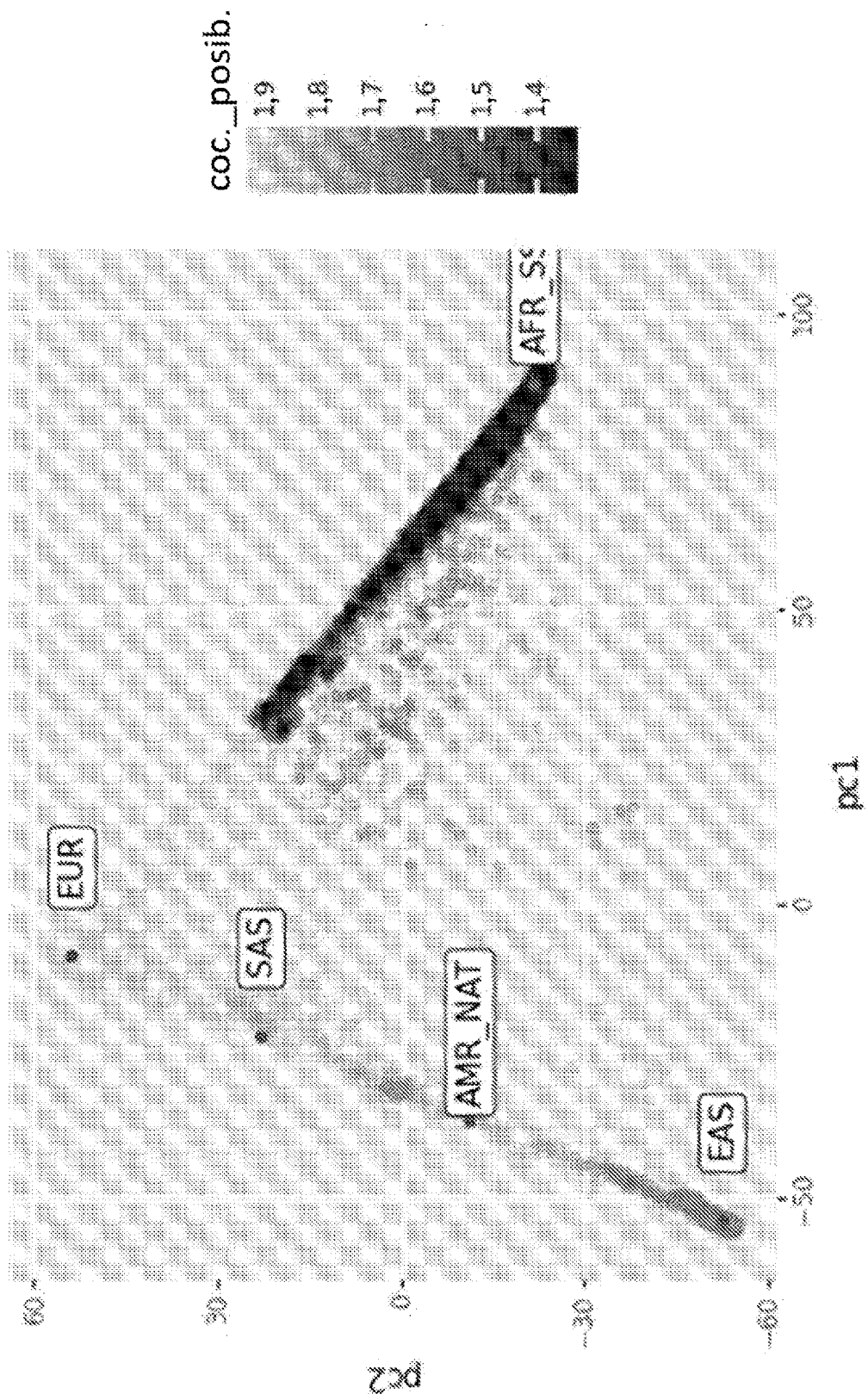


Fig. 1

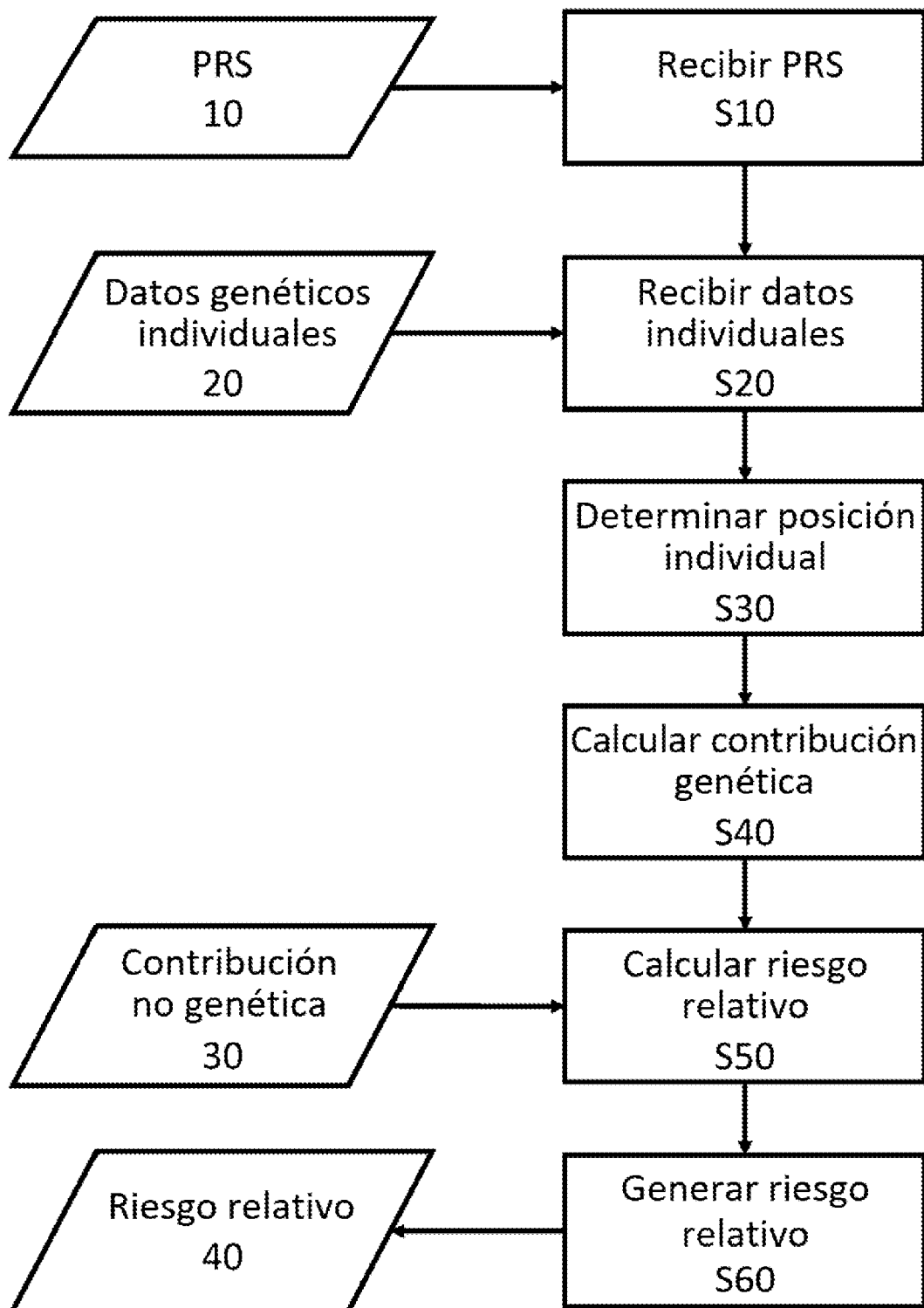


Fig. 2

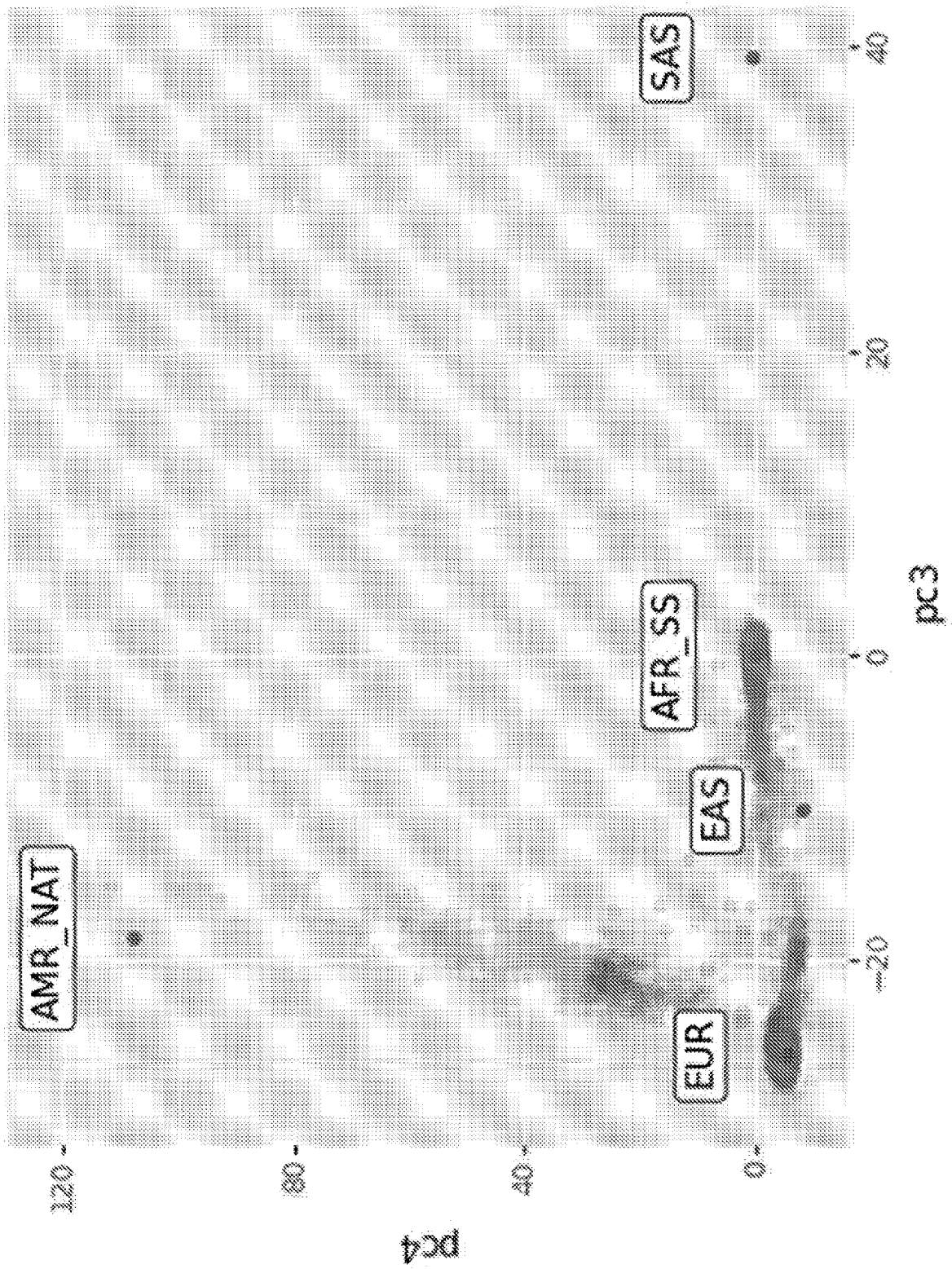


Fig. 3

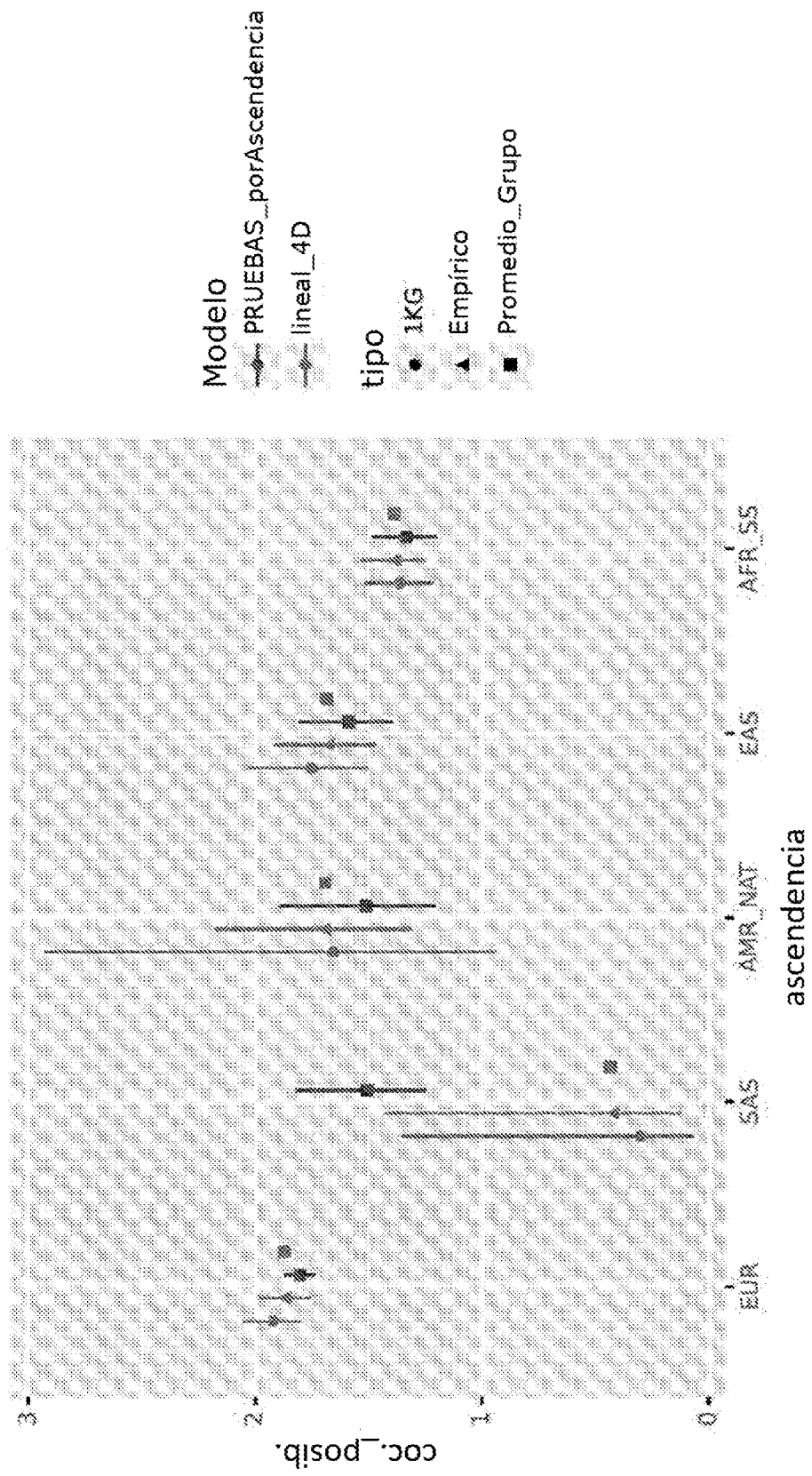


Fig. 4

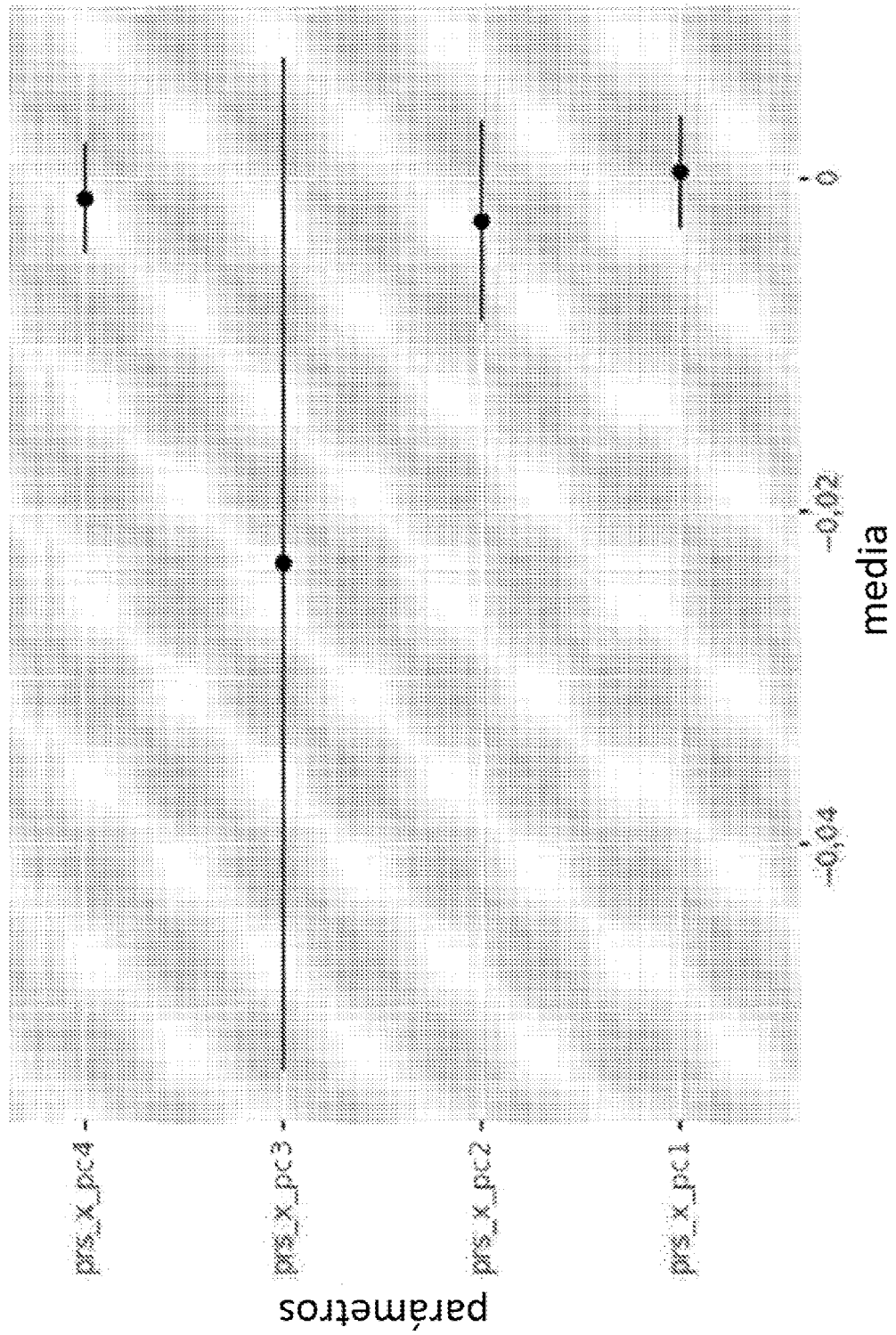


Fig. 5

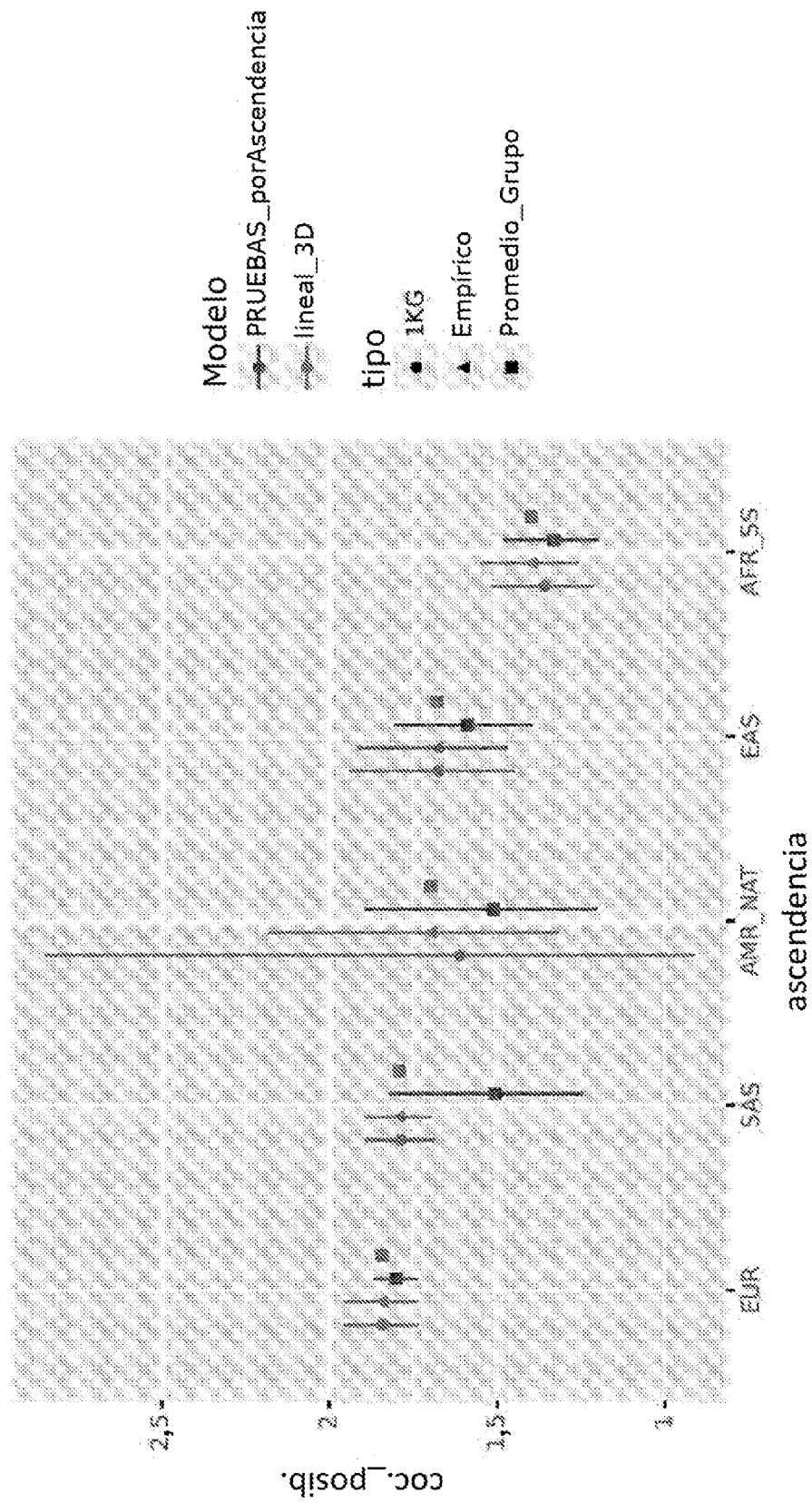


Fig. 6

Pruebas de los CP en espacio de PCA

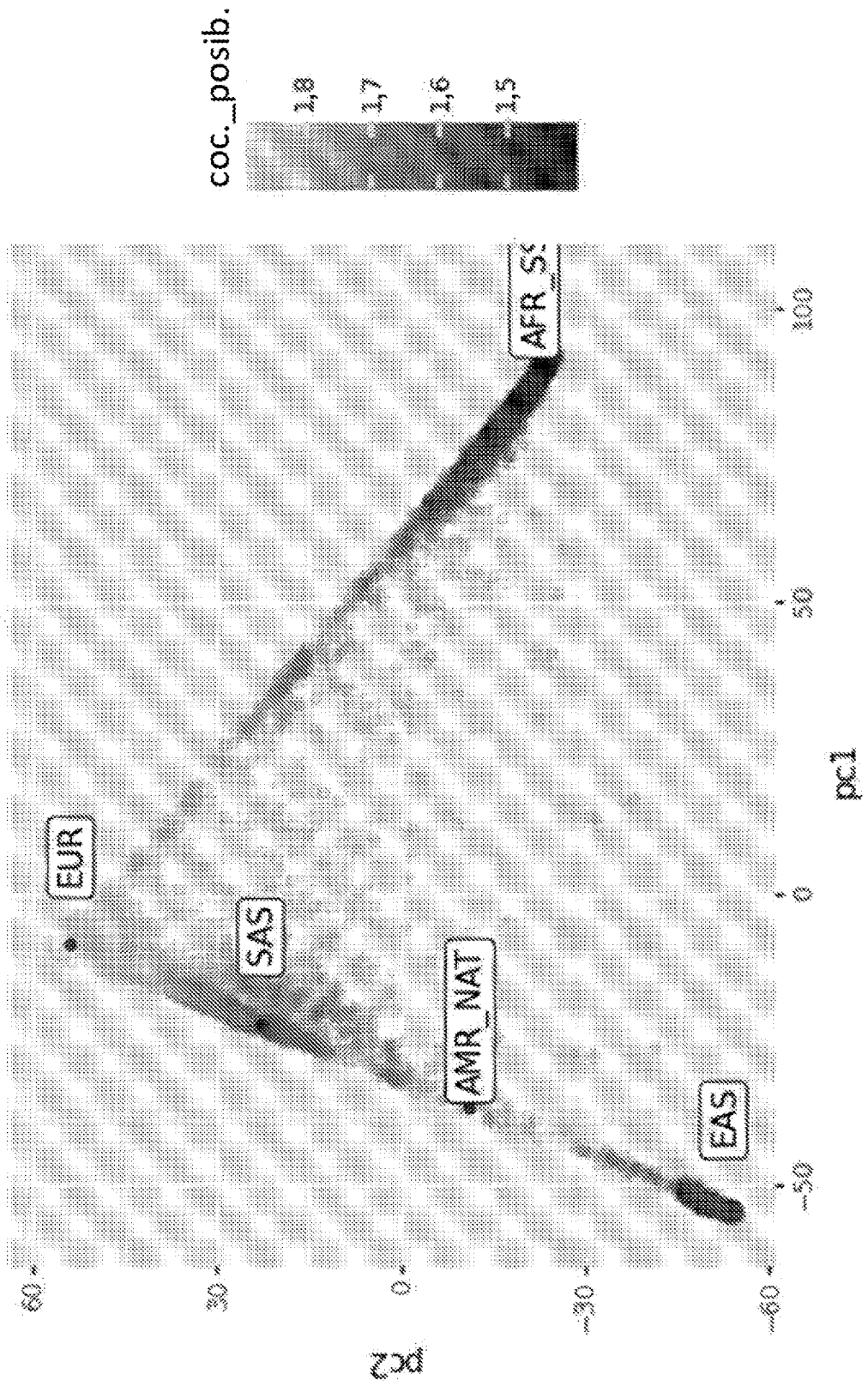


Fig. 7

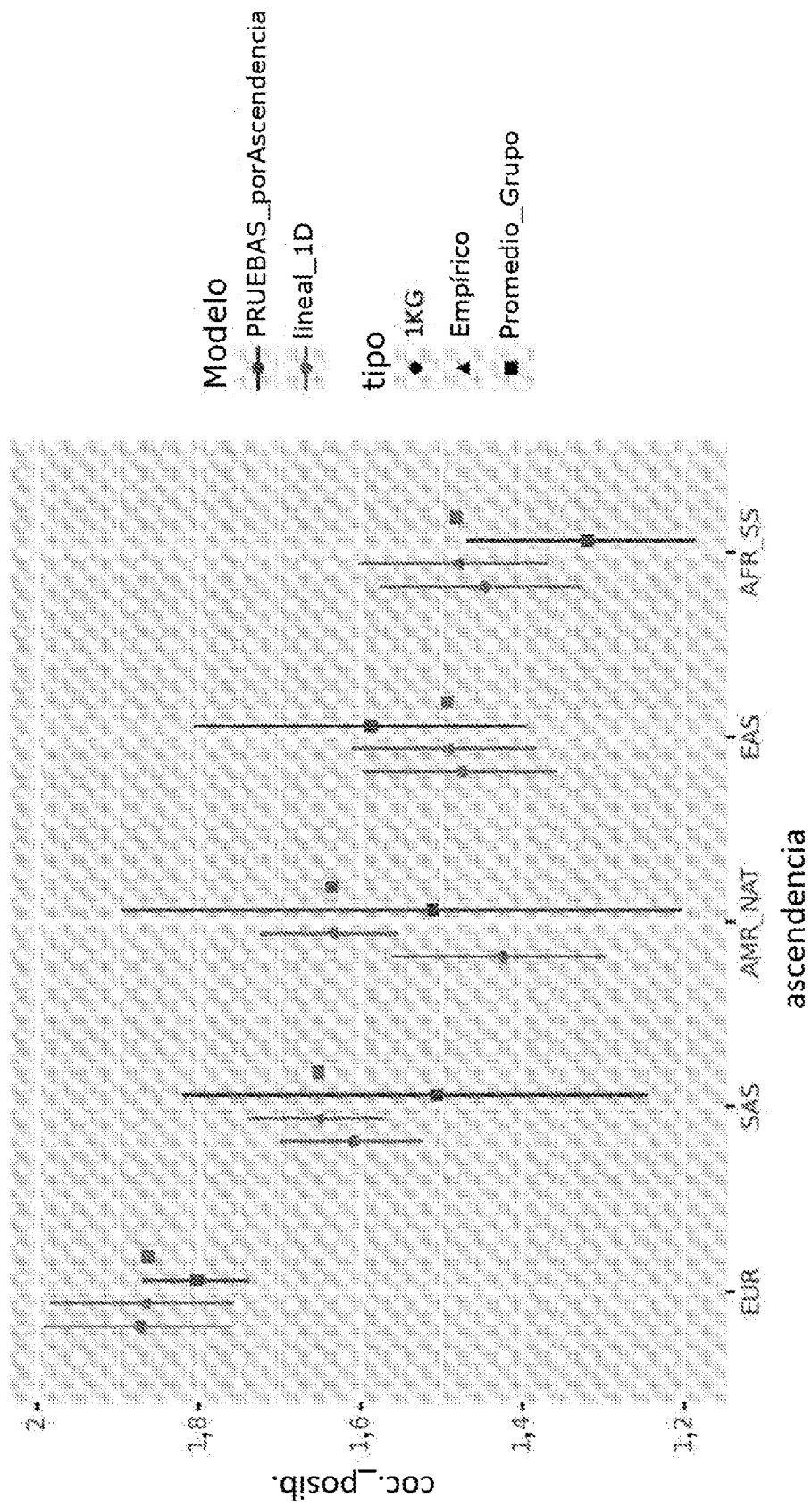


Fig. 8

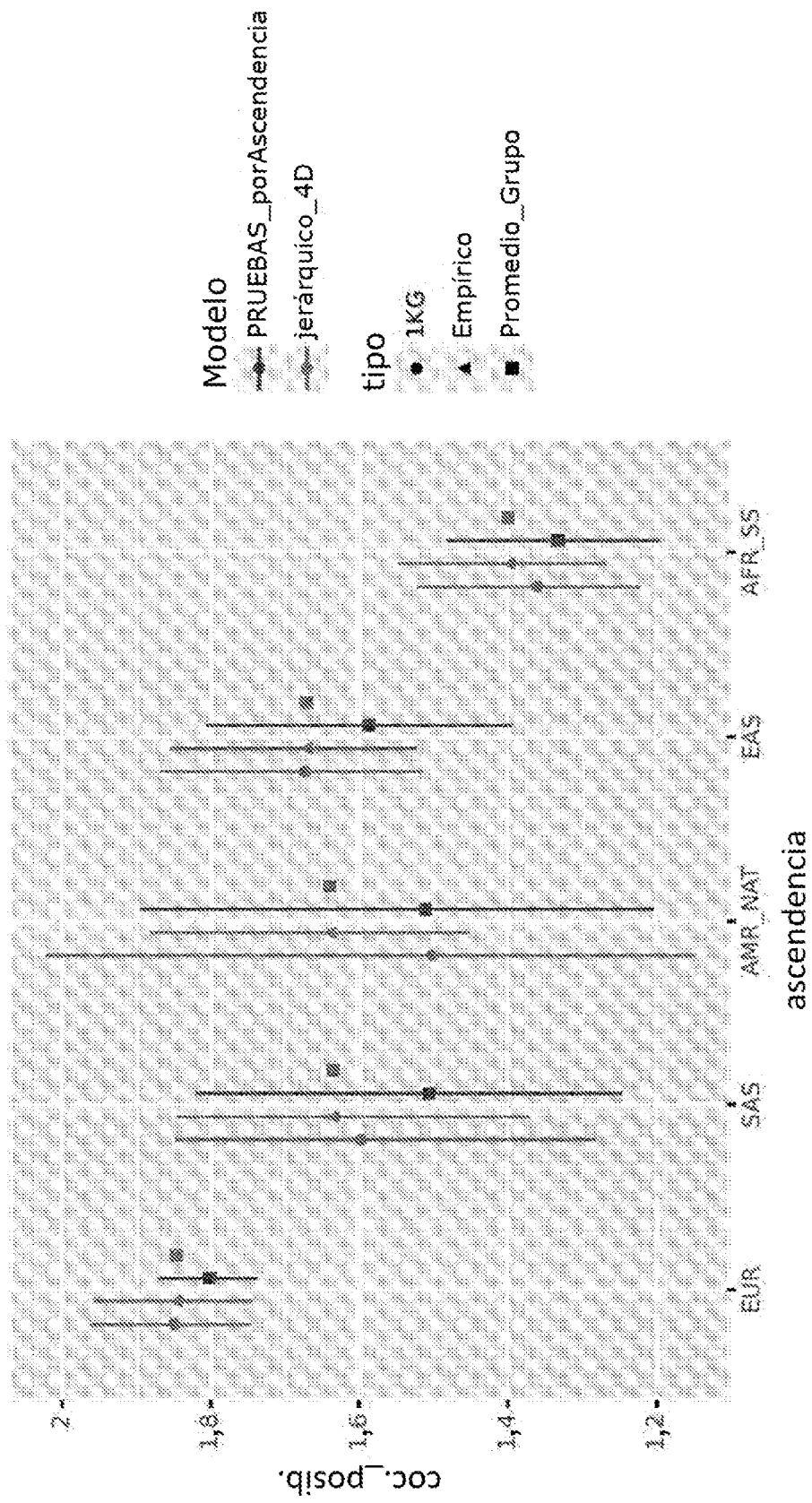


Fig. 9

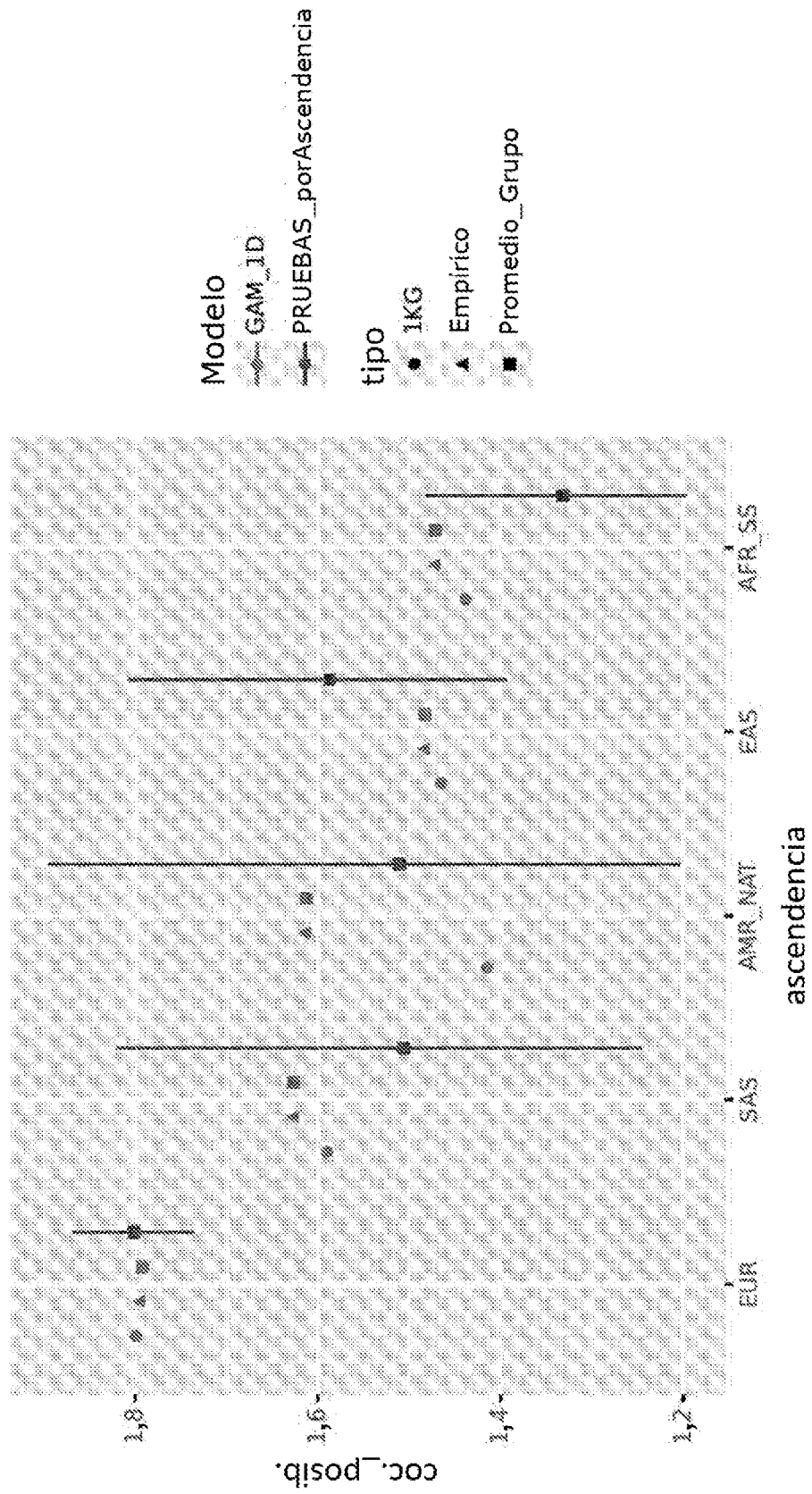


Fig. 10

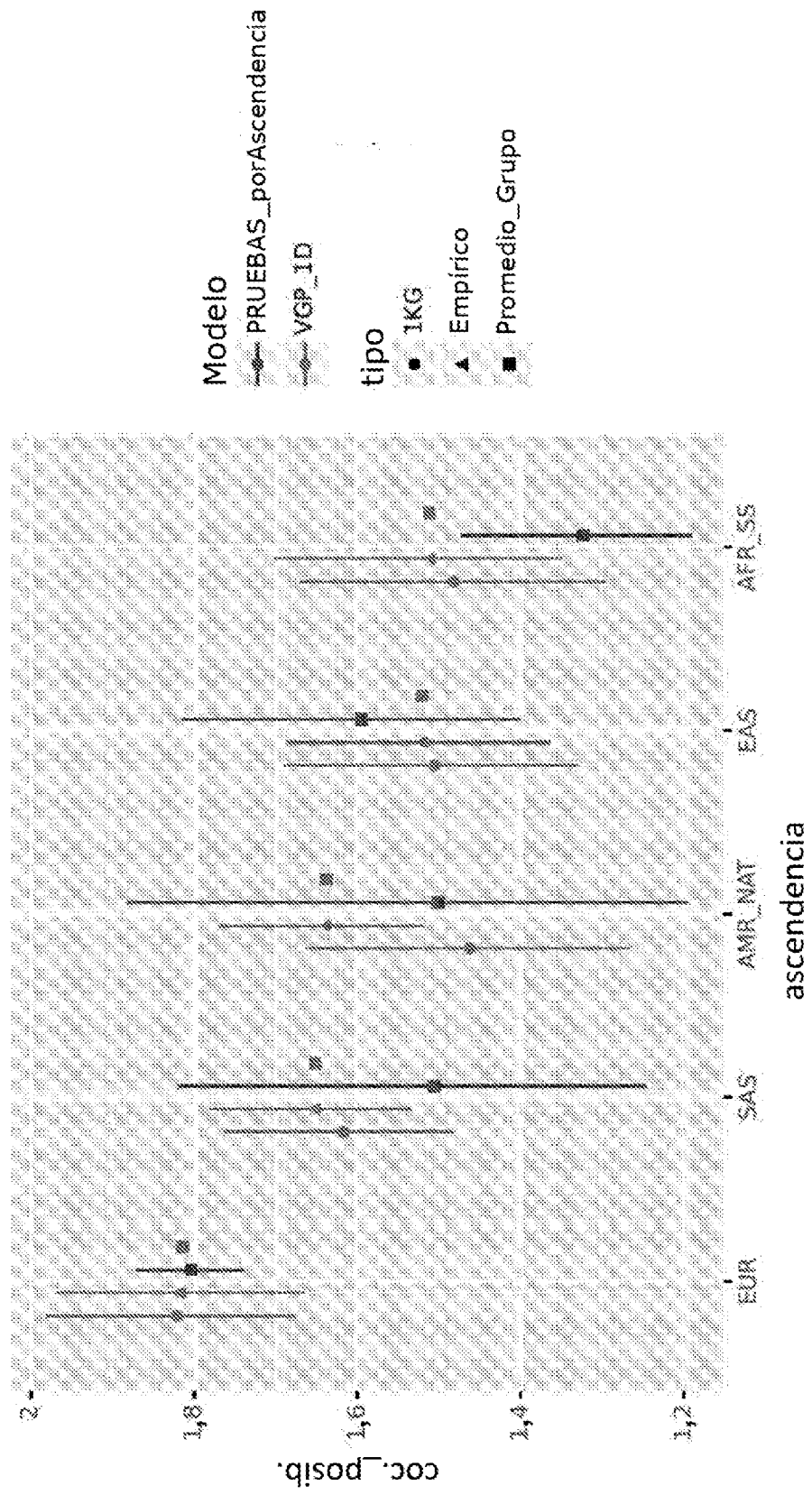


Fig. 11