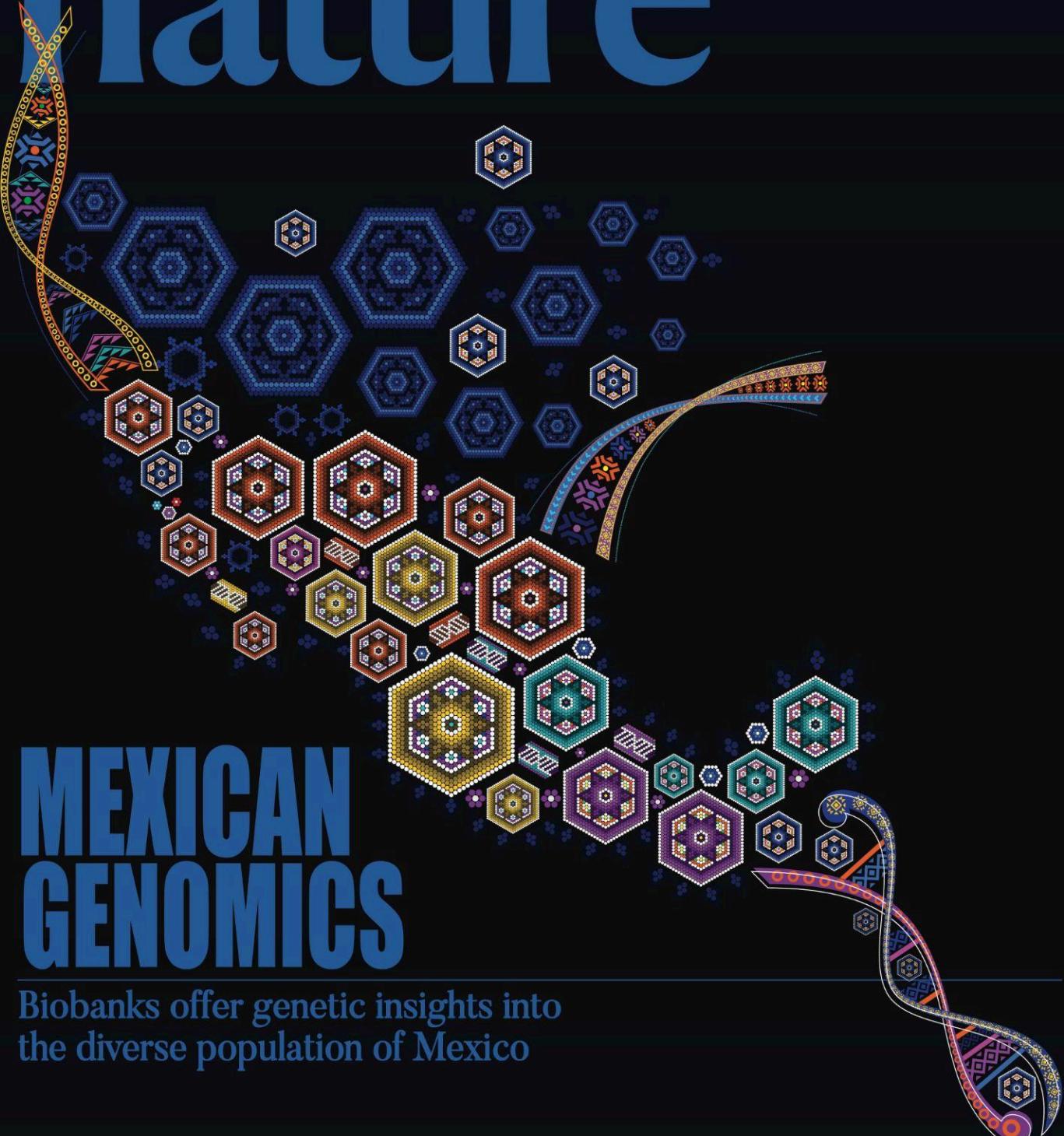


nature



MEXICAN GENOMICS

Biobanks offer genetic insights into the diverse population of Mexico

Machine intelligence
Living guidelines for the responsible use of generative AI

Heart of Mars
Molten mantle rock surrounds red planet's liquid iron core

Mutant moments
Tool to predict viral evolution could help anticipate pandemics

El Biobanco mexicano avanza la genómica médica y de poblaciones de diversas ancestrías

Mashaal Sohail^{1,2,16}✉, María J. Palma-Martínez^{1,19}, Amanda Y. Chong^{3,19}, Consuelo D. Quinto-Cortés^{1,19}, Carmina Barberena-Jonas¹, Santiago G. Medina-Muñoz¹, Aaron Ragsdale^{1,17}, Guadalupe Delgado-Sánchez⁴, Luis Pablo Cruz-Hervert^{4,5}, Leticia Ferreyra-Reyes⁴, Elizabeth Ferreira-Guerrero⁴, Norma Mongua-Rodríguez⁴, Sergio Canizales-Quintero⁴, Andrés Jimenez-Kaufmann⁴, Hortensia Moreno-Macías^{6,7}, Carlos A. Aguilar-Salinas⁴, Kathryn Auckland³, Adrián Cortés⁹, Víctor Acuña-Alonso¹⁰, Christopher R. Gignoux¹¹, Genevieve L. Wojcik¹², Alexander G. Ioannidis¹³, Selene L. Fernández-Valverde^{1,18}, Adrian V. S. Hill^{3,14}, María Teresa Tusié-Luna⁶, Alexander J. Mentzer^{3,9}✉, John Novembre^{2,15}, Lourdes García-García^{4,20}✉ & Andrés Moreno-Estrada^{1,20}✉

Publicado el 11 de octubre del 2023

Fuente original del texto y figuras:

Nature, volumen 622, páginas 775–783 (2023)

Traducción al español por Dian Barberena-Jonas

[Link al artículo en idioma original](#)

Resumen

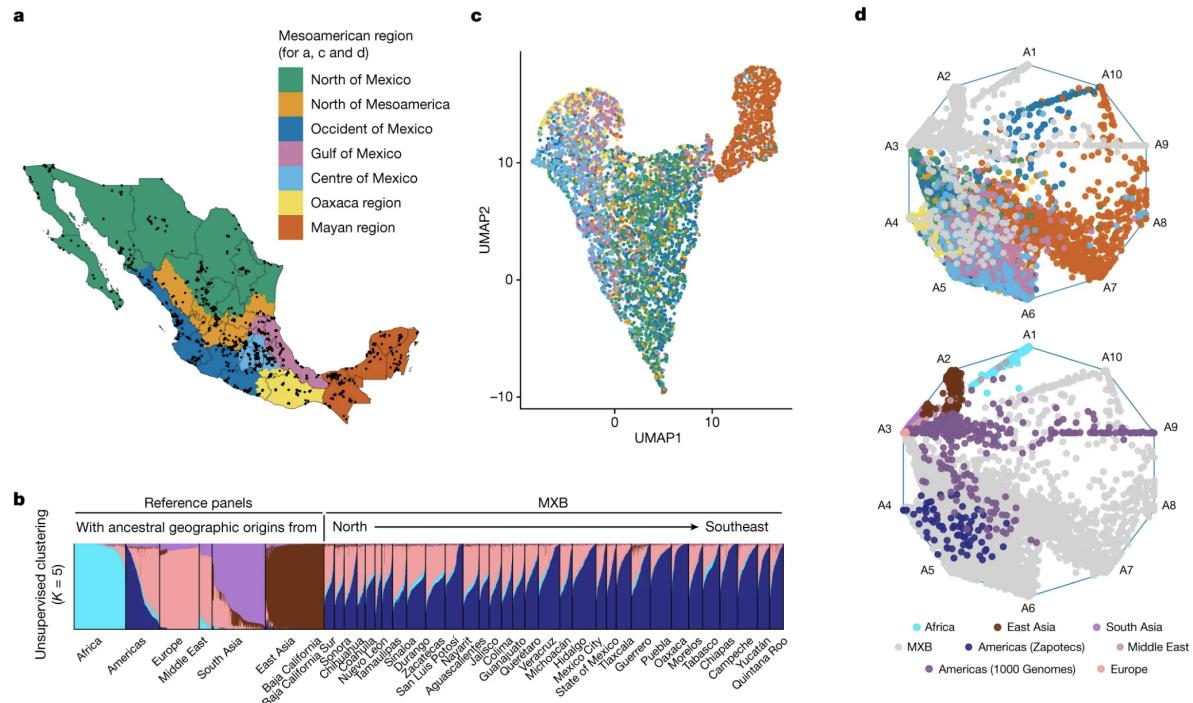
Latinoamérica continúa teniendo una representación considerablemente baja en la investigación genómica, así que las historias detalladas o las complejas características arquitectónicas del genoma se mantienen escondidas debido a la falta de datos¹. Para llenar este vacío, el proyecto del Biobanco Mexicano genotipó 6.057 individuos de 898 localidades rurales y urbanas en los 32 estados de México con una resolución de 1,8 millones de marcadores de todo el genoma con información vinculada de rasgos complejos y enfermedades, creando una valiosa base de datos de genotipo-fenotipo a nivel nacional. Aquí, utilizando la deconvolución de la ancestría y la inferencia de los segmentos idénticos por descendencia, inferimos los tamaños de las poblaciones ancestrales en las regiones mesoamericanas a lo largo del tiempo, desentrañando las dinámicas demográficas indígenas, coloniales y poscoloniales. Observamos una variación en las corridas de homocidad entre regiones genómicas con diferentes ancestrías que reflejan distintas historias demográficas y, a su vez, diferentes distribuciones de variantes deletéreas raras. Realizamos estudios de asociación de todo el genoma (GWAS, por sus siglas en inglés) para 22 rasgos complejos y descubrimos que varios rasgos se predicen mejor utilizando el GWAS del Biobanco Mexicano en comparación con el GWAS del Biobanco del Reino Unido. Identificamos factores genéticos y ambientales asociados con la variación del rasgo, como la longitud del genoma en corridas de homocigosidad como predictor del índice de masa corporal, los triglicéridos, la glucosa y la altura. Este estudio proporciona información sobre las historias genéticas de los individuos en México y disecciona sus complejas arquitecturas de rasgos, ambas cruciales para hacer que las iniciativas de medicina de precisión y preventiva sean accesibles en todo el mundo

¹Unidad de Genómica Avanzada (UGA-LANGEBIO), Centro de Investigación y Estudios Avanzados del IPN (Cinvestav), Irapuato, Mexico. ²Department of Human Genetics, University of Chicago, Chicago, IL, USA. ³The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. ⁴Instituto Nacional de Salud Pública (INSP), Cuernavaca, Mexico. ⁵División de Estudios de Posgrado e Investigación, Facultad de Odontología, Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico. ⁶Unidad de Biología Molecular y Medicina Genómica, Instituto de Investigaciones Biomédicas UNAM/Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. ⁷Universidad Autónoma Metropolitana, Mexico City, Mexico. ⁸División de Nutrición, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. ⁹Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. ¹⁰Escuela Nacional de Antropología e Historia (ENAH), Mexico City, Mexico. ¹¹Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. ¹²Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ¹³Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ¹⁴The Jenner Institute, University of Oxford, Oxford, UK. ¹⁵Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. ¹⁶Present address: Centro de Ciencias Genómicas (CCG), Universidad Nacional Autónoma de México (UNAM), Cuernavaca, Mexico. ¹⁷Present address: Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI, USA. ¹⁸Present address: School of Biotechnology and Biomolecular Sciences and the RNA Institute, The University of New South Wales, Sydney, New South Wales, Australia. ¹⁹These authors contributed equally: María J. Palma-Martínez, Amanda Y. Chong, Consuelo D. Quinto-Cortés. ²⁰These authors jointly supervised this work: Lourdes García-García and Andrés Moreno Estrada. ✉e-mail: mashaal@ccg.unam.mx; alexander.mentzer@ndm.ox.ac.uk; garcigar@insp.mx; andres.moreno@cinvestav.mx

Artículo

La arquitectura de los rasgos complejos en los humanos solo puede entenderse completamente en el contexto de la historia. El México actual abarca siete regiones culturales, incluida gran parte de Mesoamérica, con ricas historias de civilización⁹. Se han utilizado enfoques arqueológicos y antropológicos para regionalizar México en el norte de México, el norte de Mesoamérica, el centro, occidente y el Golfo de México, Oaxaca (refiriéndose aquí a la región cultural de Oaxaca) y la región maya (Fig. 1a). Estas regiones se basan en civilizaciones y culturas indígenas específicas, que florecieron temprano en la región maya, Oaxaca y occidente y el Golfo de México, y más tarde en el centro y norte de Mesoamérica. Tales historias también se han utilizado para clasificar la cronología mesoamericana en períodos preclásicos, clásicos, posclásicos, coloniales y poscoloniales¹¹.

Fig. 1: Mosaico de patrones de ancestría en el MXB y la diversidad genética en México



a, Muestreo para el MXB ($n = 5,812$ individuos con valores de latitud y longitud), que muestra a México regionalizado en regiones mesoamericanas de acuerdo con un contexto antropológico y arqueológico. **b**, Agrupamiento no supervisado ADMIXTURE utilizando paneles de referencia global ($n = 9,007$ incluyendo MXB) del Proyecto 1000 Genomas, el Proyecto de Diversidad del Genoma Humano y la Arquitectura de la Población utilizando el Estudio de Genómica y Epidemiología. **c**, Análisis uniforme de aproximación y proyección múltiple (UMAP) de MXB ($n = 5,622$) coloreado por región mesoamericana. **d**, Análisis arquetípico de MXB ($n = 5,833$) con individuos globales de referencia como en **b**, coloreado por región (arriba) o en gris (abajo). Este enfoque determina la posición de cada individuo en un espacio de diez dimensiones que en esta visualización se reduce a dos

dimensiones. Los individuos de referencia (parte inferior) se colorean utilizando grupos inferidos de **ADMIXTURE de b**. Por ejemplo, para las Américas (1000 genomas) y Oriente Medio, donde se infieren múltiples clústeres, se utiliza un color que combina estos colores de clúster.

En los últimos 500 años, la colonización española ha dejado una huella indeleble en este tapiz indígena. En un contexto colonial y poscolonial, las ancestrías genéticas que se remontan principalmente a Europa occidental, África occidental y Asia oriental se pueden identificar en los mexicanos actuales^{12,13,14,15,16}. Estas ancestrías genéticas varían en estructura y tiempo entre las regiones mesoamericanas y dan lugar a una extensa subestructura de población a pequeña escala y fuentes ancestrales en todo México^{12,13,14,15,16}. Además, se ha demostrado que tales historias genéticas variables, capturadas por las distribuciones de ancestría, afectan la variación en rasgos complejos como la capacidad pulmonar¹² y una serie de otros rasgos complejos y enfermedades¹⁷.

Sin embargo, sigue habiendo una gran brecha en la representación de mexicanos de todo México en cohortes con genotipos y fenotipos vinculados. Tal representación podría permitir estudios a mayor escala de las historias genéticas y una mejor comprensión de las arquitecturas de rasgos complejos entre individuos con diversas ancestrías de las Américas y aquellos que viven en áreas rurales¹⁸. Los análisis anteriores sobre rasgos complejos se han limitado al estudio de individuos de los EE. UU. y la Ciudad de México^{12,17}. Tampoco han modelado simultáneamente la influencia en la variación de rasgos complejos de una amplia gama de factores genéticos y ambientales como es posible con un biobanco nacional.

Para cerrar esta brecha, lanzamos el proyecto del Biobanco Mexicano (MXB, por sus siglas en inglés), que genotipa densamente a 6,057 individuos de 898 localidades distribuidas en todo el país (Figs. 1 y 2) reclutados por el Instituto Nacional de Salud Pública en los 32 estados de México. Para seleccionar las muestras para la caracterización genómica y bioquímica, enriquecimos para aquellos individuos que hablan una lengua indígena al tiempo que maximizamos la cobertura geográfica y la inclusión de localidades rurales (alrededor del 70% del MXB; Figuras 2–5 suplementarias). De los participantes en el MXB, el 70% son mujeres y comprende datos de individuos nacidos entre 1910 y 1980 (Tabla suplementaria 1) que fueron genotipados en aproximadamente 1,8 millones de polimorfismos de un solo nucleótido (SNP) y tienen información vinculada para rasgos complejos, marcadores socioculturales y biogeográficos (Tabla Suplementaria 2).

Aquí, aprovechamos la rica información arqueológica y antropológica para guiar un análisis regionalizado de México y aprovechamos el poder de la estimación de ancestría local de todo el genoma y los segmentos idénticos por descendencia (IBD) para descifrar historias genéticas a escala fina utilizando enfoques específicos de ancestría para denotar orígenes y cambios históricos en el tamaño de la población^{4,19}. Revelamos un panorama muy heterogéneo de ambos, pintando una imagen genéticamente informada de las diferentes trayectorias demográficas en las regiones mesoamericanas, incluidas las migraciones y

dinámicas coloniales. Investigamos más a fondo el papel de estas historias evolutivas capturadas por representantes de ancestrías genéticas en la configuración de la variación genética y los patrones de rasgos complejos en el México actual. Mostramos que estas historias dan como resultado patrones marcados geográficos y específicos de ancestría en las distribuciones de corridas de homocigosidad (ROH, por sus siglas en inglés) y de la carga genómica de variantes deletéreas raras. Llevamos a cabo análisis GWAS en 22 rasgos binarios y cuantitativos, y comparamos el rendimiento de predicción de puntuaciones poligénicas calculadas utilizando nuestros datos GWAS o los GWAS del UK Biobank (Bio banco del Reino Unido por sus siglas en inglés). Por último, dado que las historias evolutivas (capturadas por ancestrías genéticas) podrían asociar genotipos específicos de rasgos relevantes con ciertos antecedentes genéticos, estudiamos el impacto de las ancestrías genéticas, porciones del genoma en ROH, puntuaciones poligénicas y otros factores socioculturales y biogeográficos en la creación de variación en rasgos complejos y médicalemente relevantes en México.

Diversas ancestrías a través de escalas de tiempo

Comenzamos analizando la estructura de la población en el MXB en diferentes resoluciones geográficas y escalas de tiempo (ver la sección titulada ‘Nota sobre ancestría genéticas’ en los Métodos; Fig. 1 y Figs. 6–24 suplementarias). Dada la historia de México, en la que se espera que los linajes genéticos se remonten a regiones geográficas dispares (por ejemplo, América, Europa Occidental, África Occidental y Asia Oriental) en los últimos aproximadamente 500 años, primero analizamos a cada individuo en un marco que infiere proporciones de ancestrías genéticas sobre la base de la similitud genética con otros individuos (utilizando ADMIXTURE²⁰) en una muestra de referencia global. Utilizamos un enfoque similar para etiquetar segmentos locales en los genomas de los individuos del estudio. Usamos el término ‘ancestrías de las Américas’ cuando nos referimos a ancestrías genéticas que derivan de ancestros genéticos que vivían en las Américas antes de la colonización europea; estos también han sido referidos como ancestrías indígenas, y en algunos lugares a continuación también usamos este término (Fig. 1b, Figs. 11 y 12 suplementarias y Tabla suplementaria 3).

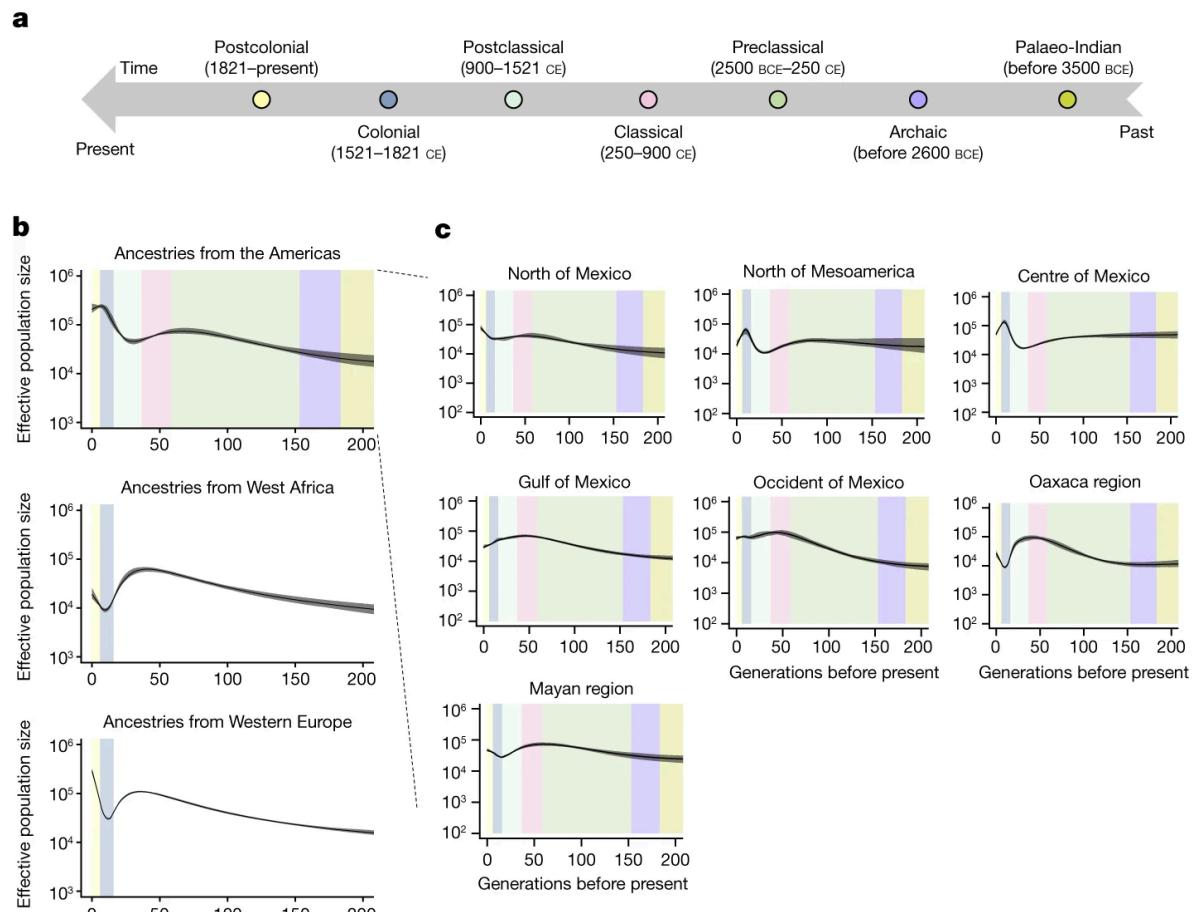
Las proporciones de ancestrías más altas de las Américas se infieren en los estados del centro y sur de México, en comparación con los estados del norte, y las ancestrías de África occidental se observan en todos los estados²¹ (Tabla suplementaria 3) consistente con los registros históricos de los viajes marítimos de la trata transatlántica de esclavos^{21,22} (Fig. 14). Observamos la presencia de una proporción pequeña pero sustancial de ancestría del este de Asia en casi todos los estados (0–2.3%), la más alta en el estado de Guerrero (2.3%), y una proporción aún más modesta de ancestrías del sur de Asia en la mayoría de los estados también (0–0.8%). Estos probablemente reflejan las migraciones de Asia a México que datan de la ruta comercial del Galeón de Manila en los siglos XVI y XVII^{1,6,23,24,25,26} y más tarde las migraciones de los siglos XIX y XX desde China y Japón, especialmente al norte de México^{27,28,29}.

Observamos la diferenciación genética más significativa a lo largo de una clina de norte a sureste en México (medida utilizando FST, que es un índice que cuantifica la proporción de la varianza genética total contenida en las subpoblaciones (S) en relación con la varianza genética total (T); Figuras suplementarias 15–18). Al considerar los autosomas de solo individuos con una proporción $\geq 90\%$ de ancestría de las Américas (inferidos mediante ADMIXTURE), la región maya de Chiapas, Tabasco, Yucatán, Quintana Roo y Campeche muestran valores de FST relativamente mayores con las otras regiones (Figs. 17 y 18 suplementarias). Esta distinción también es evidente utilizando grupos ancestrales inferidos por ADMIXTURE (Fig. 13 suplementaria) y técnicas de reducción de dimensionalidad que destacan esta subestructura de población dentro de México (Fig. 1c,d y Figs. 7–20 suplementarias). Los individuos de la región maya tienden a agruparse en su mayoría, pero se superponen con los individuos del Golfo de México y el centro de México, de acuerdo con las historias orales. En el resto de las regiones, la subestructura sutil que refleja la geografía mesoamericana es visible en el MXB, probablemente reflejando tanto las historias demográficas locales únicas de las ancestrías indígenas como los efectos del movimiento y la mezcla entre las diferentes regiones. En comparación con los muestreos y análisis anteriores que se centraron en grupos indígenas con diversos grados de aislamiento en México¹², el MXB revela niveles promedio más bajos de FST y subestructura, probablemente debido al muestreo más amplio (aunque la subestructura presentada por la región maya es más evidente en el MXB). El método de la ref. 5 destaca además la diversidad ancestral reflejada por las muestras de MXB que se representan como mezclas de múltiples fuentes (Fig. 22 suplementaria) en presencia de referencias globales (Fig. 1d y Figs. 21–23 suplementarias). Los individuos de la misma región (por ejemplo, la región maya) se modelan como mezclas de varias fuentes, lo que refleja la diversidad de la variación de ancestría dentro de esta y otras regiones mesoamericanas. Dada esta variación entre las ancestrías de las Américas y el poder único otorgado por el MXB para explorar su impacto en la variación de rasgos complejos, también obtenemos un eje de variación dentro de las ancestrías de las Américas (Fig. 24 suplementaria y Tabla 4 suplementaria).

Inferencias sobre historias genéticas dentro de México

Los mexicanos contemporáneos derivan sus ancestrías predominantemente de diversos linajes que se encuentran en América, Europa Occidental y África Occidental. Estas fuentes ancestrales tienen diferentes historias demográficas antes de su llegada al México actual y probablemente después de su llegada a diferentes regiones mesoamericanas. Para revelar la historia de los tamaños efectivos de población (N_e) de estas tres ancestrías en el MXB, analizamos los segmentos de IBD^{4,30} estratificados por inferencia de ancestría local para cada región mesoamericana 4 (Fig. 2).

Fig. 2: Valores del tamaño efectivo poblacional (N_e) a través de ancestrías y geografías revelan las historias presentes dentro México.



a, Cronología mesoamericana que colorea diferentes períodos en la historia mesoamericana utilizando un contexto antropológico y arqueológico. **b**, Cambios en el tamaño efectivo de poblacional (N_e) específico de la ancestría cambia en las últimas 200 generaciones en México ($n = 5,436$) inferido utilizando segmentos de IBD, coloreados por cronología de **a**, asumiendo una generación por cada 30 años. **c**, Cambios en el tamaño efectivo de la población (N_e) específico de la ancestría en el tiempo para las ancestrías de las Américas en diferentes regiones de México (ver Figs. 25–29 suplementarias para otros intervalos de generación y ancestrías). $n = 1,177, 640, 952, 590, 820, 315$ y 938 para el norte de México, el norte de Mesoamérica, el centro de México, Golfo de México, el occidente de México, la región de Oaxaca y la región maya, respectivamente.

Observamos una variación a pequeña escala en las trayectorias de N_e para los linajes indígenas que interpretamos en el contexto de las diferentes historias culturales de las regiones mesoamericanas 9 (Fig. 2). Como el tiempo generacional puede variar, presentamos nuestro análisis en dos extremos de 20 y 30 años por generación³¹ (Figs. 25 suplementarias y

2c, respectivamente). Cronológicamente hablando, los arqueólogos documentan que las civilizaciones mesoamericanas florecieron primero en la región maya, en Oaxaca, en Occidente y en el Golfo de México. En estas regiones, observamos grandes Ne ya en el periodo clásico (250–900 d. C.)³². Por ejemplo, en el Golfo, donde observamos una alta Ne desde el periodo preclásico (2500 a.C.-250 d.C.), existe evidencia arqueológica, entre una miríada de otros grupos, de los olmecas en el periodo preclásico, los totonacos en el periodo clásico y los huastecos en el periodo posclásico (900–1521 d.C.)³³. En Oaxaca, observamos que Ne creció rápidamente en el período preclásico al clásico, en línea con las inferencias arqueológicas de que los zapotecas ya estaban comenzando a crear asentamientos sedentarios en el período preclásico, seguidos de un aumento en las estructuras sociales y políticas en el período clásico. El período posclásico posterior se caracterizó por el militarismo y la guerra³⁴, y nuestra evidencia genética sugiere una disminución de la población hacia el final del período posclásico. En la península de Yucatán, los mayas tuvieron una prominente expansión civilizatoria en el periodo clásico (mayor Ne observado). Comenzaron a pasar por un lento declive solo en el período posclásico debido a lo que los arqueólogos han inferido como una combinación de diferentes factores políticos y ecológicos, y esta trayectoria se apoya en la tendencia de Ne³².

Estos patrones contrastan con los del centro y norte de Mesoamérica, donde el imperio azteca tuvo un bastión más recientemente; allí vemos un aumento de Ne en el posclásico justo antes de la llegada de los españoles y en parte del período colonial, después del cual comenzamos a ver una disminución de la población en Ne. La disminución de Ne tras la llegada de los españoles es más destacada en el centro y norte de Mesoamérica. En Oaxaca y la región maya, donde las ancestrías indígenas de las Américas son más prevalentes hoy en día, como lo demuestra el análisis de ADMIXTURE(Tabla suplementaria 3), la disminución de Ne es seguida por un aumento en el período poscolonial.

Al mismo tiempo, observamos que las ancestrías de Europa occidental que entraron en el acervo genético mexicano contemporáneo experimentaron una fuerte disminución en el tamaño efectivo de la población durante el período colonial. La extensión del efecto fundador varió según la región, con el efecto más fuerte visto en Oaxaca y la región maya (Figs. 26 y 27 suplementarias)

Los antepasados de África Occidental en México revelaron efectos fundadores más fuertes que variaban según la región, con Ne que oscilaba entre 10^3 y 10^4 en el período colonial. El tamaño de la población en el período poscolonial continuó creciendo en algunas regiones como el occidente y norte de México y la región maya, en comparación con otras (Fig. 28 y 29 suplementarias). De acuerdo con los resultados anteriores sobre grupos indígenas autoidentificados^{13,14}, nuestros resultados sobre los individuos MXB resaltan la heterogeneidad de las historias grupales en las regiones mesoamericanas, así como la expansión de los linajes indígenas en el período poscolonial en varias regiones.

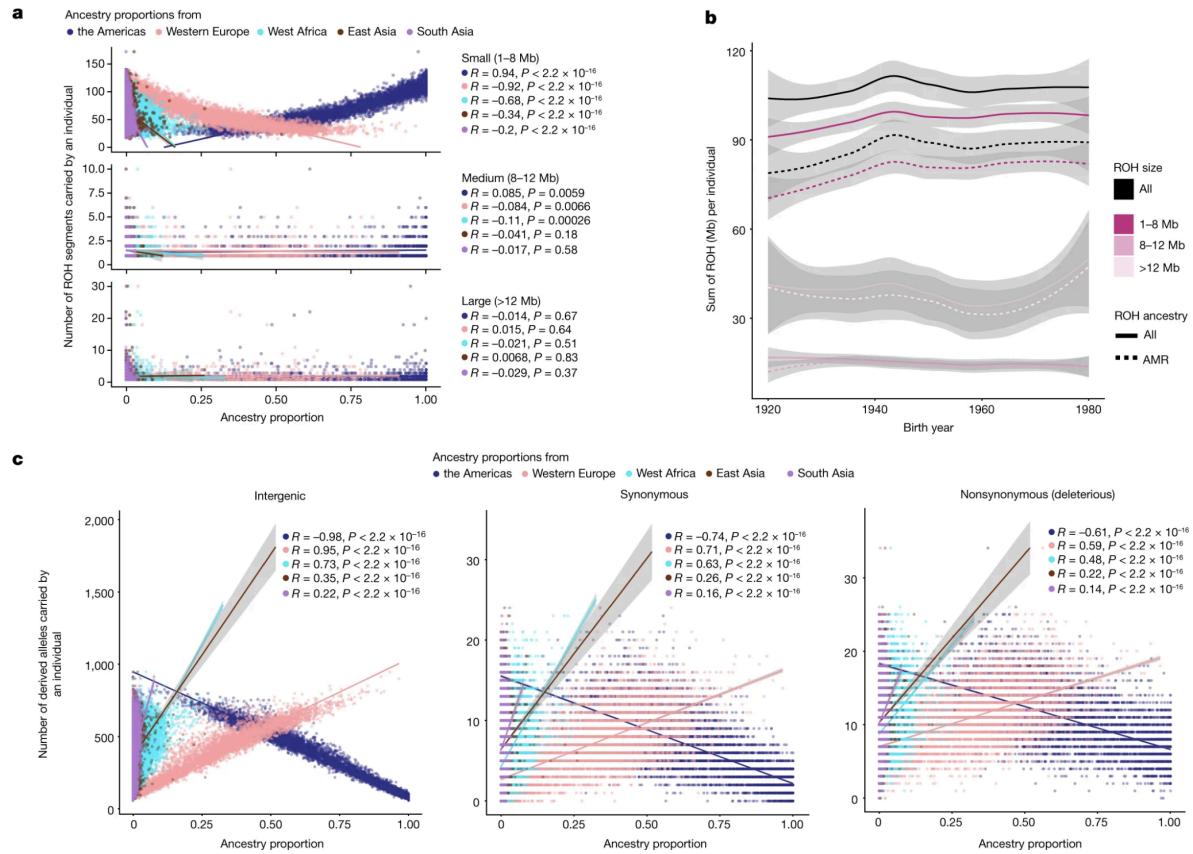
Además, generamos "gráficos de mezcla"⁶ de las regiones mesoamericanas para investigar su historia compartida utilizando un enfoque específico de ancestría y limitando el análisis a segmentos genómicos con ancestrías de las Américas. El enfoque del gráfico de mezcla

modela las diferentes regiones mesoamericanas como poblaciones en una progresión de divisiones (Datos extendidos Fig. 1a y Fig. 30 suplementaria), proporcionando información sobre las relaciones genéticas entre las diferentes regiones. Podemos observar una clara progresión de divisiones entre las poblaciones de norte a sur, con el norte de México dividiéndose primero, seguido por la ancestría común del norte de Mesoamérica y el oeste de México, seguido por la ancestría común de las regiones restantes. En particular, el centro de México y la región maya están relacionados, en consonancia con las sugerencias anteriores basadas en la IBD¹² y nuestros resultados de estructura poblacional, y ambos comparten una fuente ancestral común con Oaxaca y el Golfo de México. Estos resultados fortalecen aún más la evidencia de un corredor costero atlántico de flujo génico entre la península de Yucatán y el centro de México y el Golfo de México previamente postulado en la ref.¹². Como las historias demográficas pueden afectar los patrones de variación genética, como las distribuciones de ROH y de la carga genómica de las variantes deletéreas, a continuación evaluamos estas métricas.

Impacto de las historias genéticas en las variaciones

Analizamos los patrones de ROH en el MXB, incluida la forma en que varían a través de la geografía y los proxies de ancestría genética (inferidos de ADMIXTURE). Los patrones de ROH ayudan a iluminar aún más las historias demográficas y de la mezcla de los mexicanos³⁵, y son especialmente relevantes para la variación en rasgos complejos cuando la variación relevante para el rasgo se ve afectada por alelos de acción parcialmente recesiva³⁶. Identificamos ROH (≥ 1 Mb) en el MXB y observamos que tanto el número de ROH como la longitud total de ROH por individuo aumentan a medida que nos movemos de norte a sureste en el país (Fig. 31 suplementaria). Confirmamos que esto se debe principalmente a que los individuos con una mayor proporción inferida de ancestría genética de las Américas también tienen más ROH particularmente ROH pequeños (más pequeño que los esperados de la consanguinidad reciente; por ejemplo, <8 Mb), en sus genomas (Fig. 3a, Figs. suplementarias 32 y 33 y Tabla suplementaria 5). La aparición de muchos ROH pequeños indica coalescencias que ocurren en un período en el pasado más distante; por ejemplo, debido a un cuello de botella antiguo o un tamaño de población histórico relativamente pequeño³⁷.

Fig. 3: Efecto de patrones de variación genética en historias demográficas dentro de México



a, La prevalencia de ROH pequeños se correlaciona con los proxies de ancestría inferidos por ADMIXTURE que refleja un cuello de botella antiguo o un tamaño de población relativamente pequeño en el pasado ($n = 5,833$ individuos). **b**, Suma de ROH por individuo en función del año de nacimiento ($n = 5,833$ individuos). Las líneas continuas muestran ROH en general, y las líneas discontinuas indican que ROH se superpone a ancestrías de las Américas (AMR, por sus siglas en inglés). Los ROH se dividen en ROH pequeños, medianos y grandes, como en a. Las líneas medias condicionales suavizadas se muestran utilizando el método de suavizado del diagrama de dispersión estimado localmente. Las bandas de error representan intervalos de confianza del 95%. **c**, La carga mutacional en diferentes ancestrías muestra los efectos de los eventos de cuello de botella en la pérdida de variantes raras ($n = 5,818$ individuos). Las variantes raras se correlacionan con los niveles de ancestrías de las Américas, Europa Occidental o África Occidental (frecuencia de alelos derivados $\leq 5\%$). Las líneas medias condicionales suavizadas se muestran utilizando un modelo lineal. Las bandas de error representan intervalos de confianza del 95%. Se muestran los valores de correlación de Spearman (valores R y P de dos caras) para todas las ancestrías. El análisis de secuencias de genoma completo de 1000 muestras de Genomas MXL muestra que el resultado de la carga mutacional rara es sólido para determinar el sesgo del arreglo Global Multiétnica de Illumina (Figs. 39 y 40 suplementarias). Las variantes se anotaron utilizando la herramienta Variant Effect Predictor, y las variantes no

sinónimas (deletéreas) son un conjunto combinado de variantes sin sentido que se predice que son dañinas por polyphen2 junto con las variantes de empalme, ganancia de codón de paro y de pérdida.

Además, observamos que los ROH encontrados en segmentos genómicos indígenas son más frecuentes en individuos más jóvenes en comparación con individuos mayores (ρ de Spearman = 0.31, P = 0.016; Fig. 3b). Corroboramos que esta correlación con el año de nacimiento se deriva principalmente de ROH pequeño (ρ = 0.35, P = 0.006) y ROH pequeño encontrado en segmentos genómicos indígenas (ρ = 0.39, P = 0.002; Fig. 3b). El resultado se debe, al menos en parte, a que los individuos más jóvenes tienen mayores proporciones de ancestría indígena en comparación con los individuos mayores, especialmente en las localidades rurales (Figs. 34 y 35 Suplementarias), y coincide con observaciones recientes sobre ancestría y ROH realizadas en mexicoamericanos¹⁷. También confirmamos que esta observación no se debe al sesgo de muestreo (consulte la nota de determinación de muestreo en la Información suplementaria). La observación de ancestría más alta de las Américas en individuos más jóvenes en áreas rurales puede deberse a tasas de fertilidad más altas en áreas rurales o individuos con otras ancestrías que se mudan de áreas rurales a urbanas.

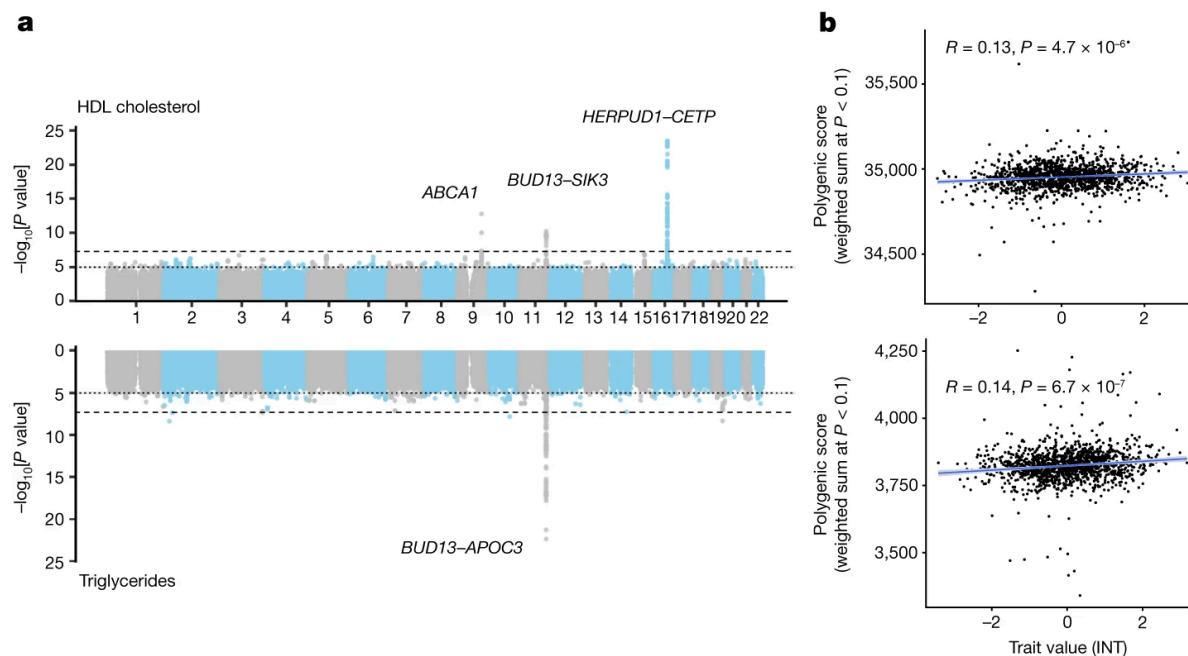
También investigamos los efectos de las historias demográficas en la distribución de frecuencias de las variantes genéticas. Este análisis está motivado por trabajos teóricos y empíricos previos que muestran que experimentar un cuello de botella cambia la distribución de frecuencia de alelos en el grupo que experimentó el cuello de botella^{38,39,40}, mientras que deja la suma general de alelos deletéreas por individuo ('carga mutacional deletéreas') sin cambios^{39,41,42}. En particular, las variantes raras se pierden o aumentan en frecuencia después del cuello de botella.

Evaluamos este efecto calculando la suma de todo el genoma de alelos intergénicos, sinónimos y supuestamente deletéreas (predichos con pérdida de sentido y con pérdida de función) por individuo. Al considerar solo alelos raros (frecuencia de alelos derivados $\leq 5\%$), observamos que los individuos con mayores proporciones de ancestría de las Américas tienen menos alelos derivados raros en todos los tipos de variantes (efecto más fuerte observado para las variantes intergénicas) (Fig. 3c) en contraste con otras ancestrías. Verificamos estas observaciones con secuencias de genoma completo de una cohorte del Proyecto 1000 Genomas (ancestría mexicana en Los Ángeles, California o MXL) (Fig. 39 suplementaria), así como con 50 genomas secuenciados como parte del proyecto MXB (Fig. 40 suplementaria), para descartar sesgos de determinación debido a la genotipificación del arreglo. Nuestro resultado probablemente refleja principalmente los efectos fundadores durante el poblamiento de América o la posterior deriva genética que condujo a la pérdida de variantes raras y/o su aumento a frecuencias más altas.

GWAS y predicción poligénica del MXB

Para comprender la transferibilidad de locus asociada a los rasgos, realizamos análisis de GWAS en 22 rasgos binarios y cuantitativos (Tabla suplementaria 6). Identificamos locus significativos para todo el genoma que pasan la corrección de Bonferroni ($P < 2.27 \times 10^{-9}$) en los cromosomas 1, 9, 11 y 16 asociados con los niveles de lípidos en la sangre (Fig. 4a). El mapeo fino de señales independientes dentro de estos locus revela variantes en o cerca de CELSR2 (lipoproteína de baja densidad (LDL): rs7528419), ABCA1 (lipoproteína de alta densidad, (HDL):rs9282541yrs2065412), el locus LINC02702-BUD13-ZPR1-APOA1-APOA4-APOA5-APOC3-SIK3,(HDL: rs180326 y rs200905431; LDL: rs66505542 ; triglicéridos: rs947989 , rs66505542 y rs5104), HERPUD1-CETP (HDL : rs57502215, rs56129100, rs193695, rs56228609 y rs117427818; colesterol: rs57502215, rs56228609 y rs118146573) y APOE (LDL: rs7412 ; triglicéridos: rs440446), que se han asociado previamente con niveles de lípidos en grupos europeos e hispanos (Tabla suplementaria 7). En particular, replicamos la asociación del alelo ABCA1*C230 que previamente se ha asociado con la disminución de los niveles de colesterol HDL ($\beta = -0.219$, s.e. = 0.030, $P = 1.64 \times 10^{-13}$; Fig. 4a), y se encuentra casi exclusivamente en grupos indígenas de las Américas ⁴³. Esta asociación se replicó en el subconjunto que contenía >90% de ancestrías indígenas inferida, aunque no alcanzó significación en todo el genoma ($\beta = -0.210$, s.e. = 0,055, $P = 1.22 \times 10^{-4}$). La restricción de la cohorte GWAS a individuos con >90% de ancestría inferida de las Américas no identificó ningún locus significativo en todo el genoma.

Fig. 4: Ejemplos ilustrativos de GWAS y predicciones poligénicas en el MXB



a, Gráficos de Manhattan que muestran los resultados de GWAS para el colesterol HDL (arriba, $n = 4,484$) y los triglicéridos (abajo, $n = 4,483$) en el conjunto de datos completo de

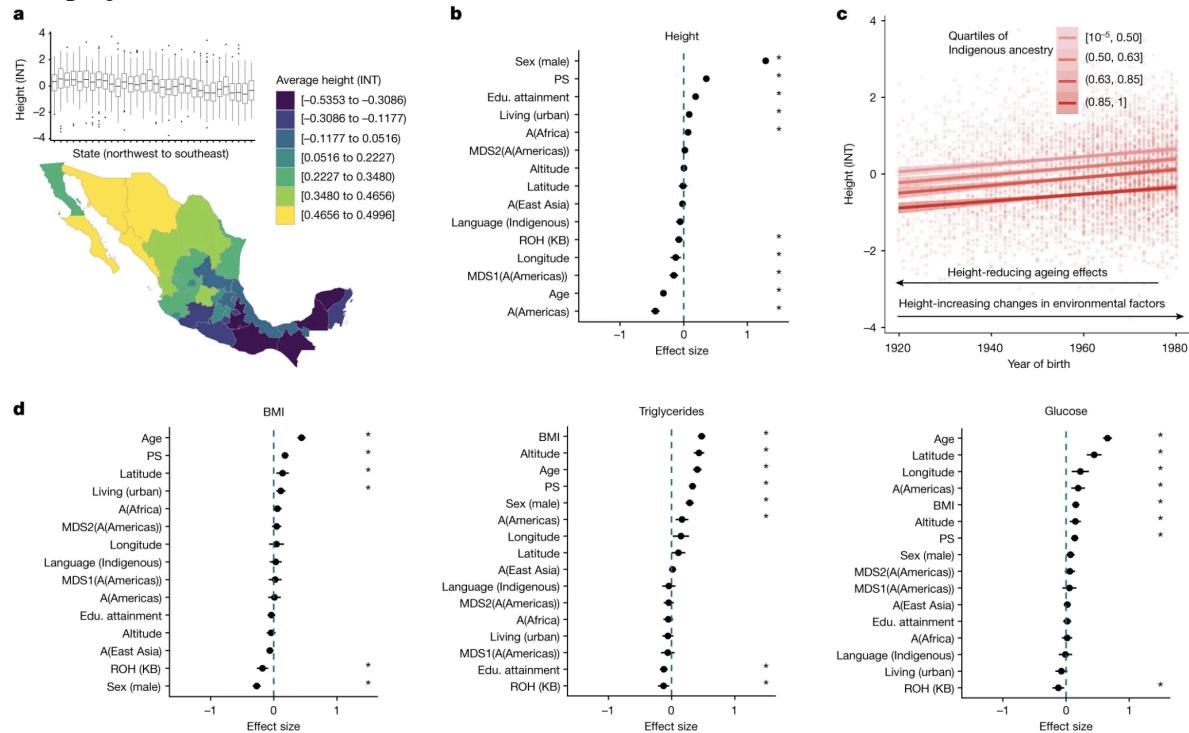
MXB. Los genes mapeados finamente están etiquetados (Métodos). Para ayudar con la visualización, se muestran 1 de cada 200 SNP con $P > 0,01$ para las gráficas de Manhattan. **b**, El rendimiento de la predicción se mide por la correlación entre la puntuación poligénica (la suma de todos los alelos asociados a $P < 0,1$ ponderada por sus tamaños de efecto estimados) y el valor del rasgo (medido por la correlación de Pearson R y su valor de P bilateral asociado) para el colesterol HDL (superior, $n = 1327$) y los triglicéridos (inferior, $n = 1326$). De acuerdo con el esquema en la Fig. 41 suplementaria, para **b**, GWAS se llevó a cabo en dos tercios del MXB, y el tercio restante del MXB se utilizó para calcular las puntuaciones poligénicas y probar su capacidad para predecir rasgos complejos. Las líneas medias condicionales suavizadas se muestran utilizando un modelo lineal. Las bandas de error representan intervalos de confianza del 95%. Las puntuaciones se calcularon utilizando genotipos de MXB imputados por TOPMed. Los rasgos se normalizaron utilizando una transformada normal inversa (INT) tanto para **a** como para **b**. Para una evaluación adicional del rendimiento de la predicción, consulte las Figuras de datos extendidos 1b y 2–10 y Tablas suplementarias 8 y 9.

Para evaluar la transferibilidad en la predicción de rasgos cuantitativos utilizando puntuaciones poligénicas, volvemos a realizar un GWAS en solo 4.000 individuos seleccionados al azar del MXB y construimos puntuaciones poligénicas en los 1.778 individuos restantes (Fig. 41 suplementaria). Calculamos las puntuaciones poligénicas utilizando datos de genotipo y genotipos imputados utilizando TOPMed. Para evaluar el impacto del uso de diferentes estadísticas de GWAS en el rendimiento de la predicción, para la comparación también calculamos puntuaciones poligénicas utilizando GWAS de panancestría del UKB a la luz de las diferentes fuentes de ancestría en México (Fig. 4b y Datos extendidos Fig. 1b). Observamos que la predicción basada en MXB funciona mejor o tan bien como la predicción basada en UKB, a pesar de un tamaño de muestra mucho menor, para la glucosa, la creatinina, el colesterol y la presión arterial diastólica (Datos Extendidos Figs. 1b y 2–10 y Tablas suplementaria 8 y 9). Los niveles de triglicéridos, HDL y colesterol LDL también son casi tan bien predichos por el MXB GWAS (Fig. 4b). Estos resultados indican que se lograrían mayores ganancias en el poder de predicción al aumentar aún más el tamaño de la muestra. Aunque es probable que muchos factores estén involucrados en la portabilidad diferencial de la puntuación poligénica por rasgo, algunas características de la arquitectura del rasgo son probablemente relevantes, como la fuerza de la selección estabilizadora bajo la cual se encuentra el rasgo, su tamaño objetivo mutacional y la heredabilidad por sitio causal^{44,45}. Utilizando los tamaños estimados mutacionales de estudios anteriores de GWAS⁴⁵, observamos que los rasgos con tamaños estimados mutacionales inferidos más pequeños (creatinina y triglicéridos) se predicen mejor con los SNP descubiertos en MXB en comparación con los rasgos que se infiere que tienen un tamaño estimados más grande (altura e índice de masa corporal (IMC))⁴⁵. Los predictores basados en UKB se utilizan en nuestro modelado de rasgos complejos a continuación, ya que se pueden calcular para todos los individuos MXB.

Rasgos arquitectónicos complejos en el MXB

Por último, evaluamos la contribución de la variación genética resultante de historias demográficas y ambientales variables o distribuciones de variantes causales que afectan la variación en rasgos complejos o enfermedades en México (Fig. 42 suplementaria). Nos centramos en varios rasgos cuantitativos: altura, IMC, triglicéridos, colesterol, glucosa, presión arterial y otros. Con el objetivo de comprender cómo se distribuyen los rasgos geográficamente y en relación con las variables de un solo modelo, primero visualizamos los valores promedio de los rasgos por unidades de nuestros factores biogeográficos y socioculturales para comprender las dimensiones de la variación de los rasgos (Fig. 5a y Figs.43–51 suplementarias).

Fig. 5: Un análisis de los factores que afectan a la altura y la variación en otros rasgos complejos.



a, Abajo: mapa de altura media en México ($n = 5,770$). La altura se normalizó utilizando un INT. Arriba: diagramas de caja de variación de altura (INT) en cada estado de noroeste a sureste. Los diagramas de caja muestran el valor medio y los cuartiles. Los bigotes se extienden a los valores mínimo y máximo. Los puntos representan valores atípicos. $n = 5,846$ muestras biológicamente independientes se utilizaron para el análisis. **b**, El modelo explicativo para la variación de la altura implica el papel de la genética y el medio ambiente. La gráfica muestra las estimaciones del tamaño del efecto y los intervalos de confianza ($1.96 \times \text{s.e.m.}$) de un análisis de modelo mixto. Todos los predictores cuantitativos están centrados y escalados en 2 desviaciones estándar. Los asteriscos

muestran significancia a una tasa de descubrimiento falso < 0.05 en todos los rasgos y predictores analizados⁵⁰. $n = 4,625$ muestras biológicamente independientes se utilizaron para el análisis. **c**, Altura en función del año de nacimiento en cuartiles de ancestría de las Américas ($n = 5,598$). Las barras de error representan intervalos de confianza del 95 %. **d**, Perfiles de rasgos para IMC (izquierda), triglicéridos (centro) y glucosa (derecha). Resultados del análisis de modelo mixto, como en **b**. La gráfica muestra las estimaciones del tamaño del efecto y los intervalos de confianza ($1.96 \times \text{s.e.m.}$) de un análisis de modelo mixto. $n = 4,607$, 3,664 y 3,613 muestras biológicamente independientes se utilizaron para el análisis de IMC, triglicéridos y glucosa, respectivamente. Para **b** y **d**, los PS son puntajes poligénicos calculados utilizando estadísticas de resumen de UKB (SNP significativos a $P < 10^{-8}$), A(África/Asia Oriental/Américas) se refiere a las proporciones de ancestría de esa región como se infiere de la ADMIXTURE, y MDS1 (A(Américas)) y MDS2(A(Américas)) se refiere a los ejes de escala multidimensional (MDS) dentro de las ancestría de las Américas como se infiere utilizando un análisis MAAS-MDS (Fig. 24 suplementaria). El nivel educativo (Edu.) está en una escala de 0 a 8 (nivel educativo bajo a alto), y la altitud se mide en metros (bajo a alto).

A continuación, utilizamos un modelo mixto para estimar la contribución de los factores genéticos a la variación de los rasgos modelados conjuntamente con los factores ambientales (Fig. 5b,d y Figs . 52–61 suplementarias). Los proxys de ancestría genética pueden asociarse con rasgos complejos debido a factores genéticos o debido a factores no genéticos que covarian con ancestría genéticas, como experiencias diferenciales de discriminación, nutrición dietética y estado socioeconómico (Fig. 42 suplementaria). Los factores genéticos que varían con los representantes de ancestría genética pueden ser diferentes distribuciones de ROH u otros patrones diferenciales de variación genética causados por historias demográficas y ambientales que varían entre ancestría. También se ha demostrado previamente que ROH tiene asociaciones con una amplia gama de rasgos complejos como la altura, el peso y el colesterol, lo que apunta a una arquitectura recesiva de estos rasgos^{36,46}. Como se muestra anteriormente, los representantes de ancestría genética en el MXB se correlacionan con el número y la longitud de ROH (Fig. 3a). Por lo tanto, desarrollamos un modelo mixto para la asociación de factores genéticos como proxys de ancestría, ROH y puntajes poligénicos con variación de rasgos. Consideramos en nuestro modelo varios factores ambientales para mejorar el poder y para cuestionar el papel de los factores genéticos reflejados en los proxies de ancestría en comparación con los factores ambientales. Incluimos variables disponibles en el MXB relacionadas con la discriminación, las oportunidades socioeconómicas y el entorno de vida (denominados colectivamente factores socioculturales y biogeográficos), así como efectos aleatorios no observados para modelar la relación críptica y los posibles factores ambientales no modelados. En este modelo, una asociación significativa con representantes de ancestría podría reflejar la asociación de genotipos causales particulares con esas ancestría o factores ambientales no modelados asociados, como la nutrición. Nuestro modelo combinado explica el 66,6% de la varianza para la altura, el 30,4% para el IMC, el 44,3% para los triglicéridos, el 30,9% para el colesterol y el 30,91%

para la glucosa.

Como ejemplo ilustrativo, los valores de altura muestran un claro patrón creciente de sureste a noroeste en el MXB (Fig. 5a). Aunque los valores de altura en cada estado exhiben una gran varianza (Fig. 5a), la altura se correlaciona significativamente con la longitud (Fig. 5b y Figs. suplementarias 43 y 45). Encontramos que las personas con una mayor proporción de ancestría indígena de las Américas son significativamente más bajas ($\rho=-0,45$, $P<2,2\times10^{-16}$) mientras que los individuos con una mayor proporción de ancestría de África Occidental son significativamente más altos ($\rho=0,07$, $P <0,005$; Fig. 5b). Además, considerando las ancestrías en una resolución más fina, observamos una disminución de la altura con un cambio en las ancestrías del norte de México (por ejemplo, Huichol y Tarahumara) a las de la región maya (por ejemplo, Tojolabal y Maya; $\rho =-0,156$, $P =6,13\times10^{-6}$; Fig. 55 suplementaria). La longitud total de ROH también se asocia significativamente con una altura más corta (0.08 ± 0.01). Simultáneamente, los individuos más jóvenes en todo el espectro de ancestría son más altos que los individuos mayores con las mismas ancestría (Fig. 5c), exhibiendo el impacto de factores no genéticos (mejora de la nutrición en el rango de años de nacimiento estudiado o efectos del envejecimiento) en la variación de la altura también.

La obesidad es un problema de salud pública en México ⁴⁷ y se ha sugerido que está relacionada con un mayor riesgo genético asociado con los ancestros indígenas ⁴⁸. Contrariamente a esta hipótesis, en el MXB en su conjunto, cuando se considera de manera univariada, la ancestría genética indígena y el habla de una lengua indígena en realidad se correlacionan con un IMC más bajo (Figs. suplementarias 49 y 60). En nuestro modelo conjunto con variables, aunque esas asociaciones desaparecen, los ROH en un genoma (que son más prevalentes en las ancestrías genéticas indígenas) también se asocian con un IMC más bajo. Por el contrario, como vivir en un entorno urbano se asocia con un IMC más alto (Fig. 5d), nuestros resultados sugieren un enfoque en factores relacionados con un entorno urbano como la dieta y el sedentarismo para ayudar a abordar el problema de la obesidad en México. Un análisis segmentado adicional que considera solo individuos en entornos urbanos sugiere lo mismo: observamos individuos que hablan una lengua indígena que se asocia con un IMC más alto solo en entornos urbanos (Fig. suplementaria 61).

Por el contrario, algunos otros rasgos muestran una correlación con la proporción de ancestros genéticos inferidos de las Américas de un individuo: creatinina ($(\beta = - 0.13, = P 0.0095)$, densidad ($\beta = - 0.141, P = 0.013$) triglicéridos ($\beta = 0.16, P = 0.001$) y nivel de glucosa en la sangre ($\beta = 0.19, P = 0.0005$; Fig. 54 suplementaria). En el MXB, la cantidad de genoma de un individuo en ROH se asocia con un IMC más bajo ($\beta = - 0.18, = P 7.11 \times 10^{-5}$), triglicéridos ($\beta = - 0.13, = P 0.004$) y nivel de glucosa en la sangre ($\beta = - 0.12, P = 0.01$; Fig. 56 suplementaria). También encontramos que las puntuaciones poligénicas calculadas utilizando SNP significativos en todo el genoma del GWAS de panancestría de UKB son un predictor significativo de la variación de rasgos complejos para todos los rasgos analizados (Fig. 57 suplementaria). La presión arterial se asocia con factores ambientales y puntuaciones poligénicas, pero no con otros factores genéticos de todo el genoma (Fig. 53

suplementaria). En particular, entre los factores ambientales investigados, vivir en un entorno urbano se asocia con niveles más altos de altura, IMC, colesterol y creatinina, mientras que vivir en altitudes elevadas se asocia significativamente con niveles más altos de triglicéridos, glucosa, colesterol, creatinina y presión arterial (Fig. 58 suplementaria). Un mayor nivel educativo se asocia con una mayor altura, LDL, HDL y niveles más bajos de triglicéridos, mientras que hablar un idioma indígena se asocia con niveles más bajos de creatinina y colesterol (Fig. 58 suplementaria).

Trabajos anteriores han implicado al alelo ABCA1*C230 (rs9282541) en la disminución de los niveles de HDL y han demostrado que este alelo es aparentemente exclusivo de las ancestrías genéticas indígenas de las Américas (que se encuentran en 29 de 36 grupos de nativos americanos, pero no en individuos europeos, asiáticos o africanos)⁴³. En el MXB, observamos de manera similar que el alelo ABCA1*C230 está en frecuencias más altas en individuos con una mayor proporción de ancestrías de las Américas, y observamos que los individuos con frecuencias de alelos ABCA1*C230 más altas tienen niveles de HDL más bajos (Fig. 59a suplementaria). Sin embargo, en general, las ancestrías indígenas no están asociadas con los niveles de HDL después de tener en cuenta otras variables. De hecho, las variantes genéticas colectivamente en el fondo genético indígena se asocian con niveles más bajos de LDL (Fig. 59 B suplementaria). Estos resultados ilustran cómo la interacción entre los factores culturales y dietéticos, y los factores genéticos son esenciales para diferentes resultados de colesterol. También implican que, aunque algunas variantes funcionales pueden ser específicas de regiones o antecedentes genéticos, estas son pocas (alrededor de 1,000 de estas variantes se estiman en las Américas con una frecuencia del 40% o más a partir del muestreo de diversos genomas del Proyecto de Diversidad del Genoma Humano⁴⁹), y advierten contra el uso de la proporción de ancestría global de un individuo como predictor del efecto de una sola variante funcional. Nuestros resultados en general respaldan que las variantes funcionales con frecuencias variables o interacciones ambientales son parcialmente responsables de la variación en una gama de rasgos complejos en México⁴³.

Conclusión

Nuestro trabajo demuestra el valor de generar datos de genotipo-fenotipo en grupos subrepresentados para revelar historias genéticas menos conocidas y generar hallazgos de relevancia biomédica. También es una ilustración del modelado en conjunto de los efectos genéticos y ambientales para revelar la etiología de rasgos complejos y enfermedades. En este proyecto, aseguramos la diversa presencia indígena y rural en nuestra estrategia de muestreo, consideramos la fluidez de las ancestrías de diferentes regiones locales y globales en nuestros análisis y evaluamos su reflejo en la variación de rasgos complejos genéticos y relevantes para la enfermedad. Al aprovechar el mayor biobanco genómico nacional en México, encontramos diversas fuentes de ancestría en México a la luz de su historia única, e inferimos historias demográficas y de mezcla y ROH utilizando la identidad de haplotipos específicos de ancestría que revelan una elaborada estructura a escala fina en el país. Observar un mayor número de ROH pequeños en individuos más jóvenes en el MXB y en segmentos genómicos de ancestría indígena es relevante para analizar la arquitectura genética de rasgos y

enfermedades complejas, especialmente aquellas con un componente recesivo. También mostramos que la historia demográfica afecta la distribución de frecuencia de las variantes genéticas, cambiando así la cantidad de variantes raras que portan los individuos con diferentes ancestrías. Demostramos el valor de los GWAS realizados sobre un recurso como el MXB para predecir rasgos complejos. El MXB GWAS muestra utilidad para el cálculo de puntajes poligénicos en cohortes mexicanas independientes, así como para el metanálisis con otras cohortes de GWAS para aumentar aún más el poder de predicción. Por último, observamos un impacto significativo de las ancestría genéticas en diferentes escalas de tiempo, ROH, puntajes poligénicos y variables socioculturales y biogeográficas en varios rasgos complejos que implican la importancia de los factores genéticos y ambientales para explicar la variación de rasgos complejos y en consideraciones de posibles intervenciones de salud pública. Nuestros resultados muestran la importancia añadida de considerar los factores genéticos para la medicina preventiva y personalizada más allá de los factores ambientales. Nuestros resultados informarán el diseño de futuros estudios de rasgos genéticos y complejos en México y América Latina. Esperamos motivar esfuerzos adicionales para fortalecer la capacidad de investigación local en toda América Latina y beneficiar a los grupos desatendidos a nivel mundial.

Métodos

Encuesta Nacional de Salud 2000

Desde 1988, México ha establecido encuestas nacionales de salud periódicas (Encuesta Nacional de Salud (ENSA), originalmente concebidas como Encuestas Nacionales de Nutrición) para la vigilancia de la nutrición y las métricas de salud basadas en la población mexicana. En este estudio, utilizamos datos y muestras recogidas de la encuesta realizada en 2000, la ENSA 2000. Esta fue una encuesta probabilística, de múltiples etapas, estratificada y por conglomerados realizada por la Secretaría de Salud de México desde noviembre de 1999 hasta junio de 2000. El diseño y los métodos de investigación se han descrito en otra parte⁵¹. Los participantes fueron seleccionados al azar para ser representativos de la población civil mexicana no institucionalizada a nivel estatal y nacional. Personal capacitado realizó las entrevistas. Se recopiló información sobre las características sociodemográficas y del hogar, el estado de salud actual, el uso de los servicios de atención médica y los aspectos conductuales de los participantes. Los sueros y las capas leucocitarias se obtuvieron de 43.085 individuos de 20 años o más. De esta encuesta han surgido más de cincuenta publicaciones que proporcionan información crítica sobre el estado de la salud nacional junto con algunos rasgos genéticos de la población muestreada⁵². En particular, la inclusión de personas de lugares remotos y rurales en México hace que esta encuesta sea única. Dado su gran volumen, diseño de muestreo sofisticado, amplitud de muestreo demográfico y amplios datos de rasgos, la ENSA 2000 representa un valioso recurso genético sin explotar para vincular los marcadores genéticos y los resultados de salud.

Fenotipo, estilo de vida y ambiente para el Proyecto MXB

Para cada individuo, tenemos acceso a una variedad de datos antropométricos, de enfermedades, de estilo de vida y ambientales. Estas variables se resumen en la Tabla suplementaria 2. Las muestras de suero se utilizaron además para medir una serie de rasgos bioquímicos analizados en este estudio. Todos los rasgos analizados en el análisis de rasgos complejos se preprocesaron de la siguiente manera.

Los datos biométricos se filtraron para eliminar valores atípicos con errores aparentes en la entrada de datos. Los valores atípicos se identificaron sobre la base de la densidad de distribución en el conjunto de datos completo de >6.000 individuos, lo que resultó en una altura entre 100 y 200 cm y un peso entre 25 y 300 kg.

Los rasgos bioquímicos se seleccionaron de manera similar para eliminar los extremos y los valores negativos (<0). La glucosa también se comparó con las pruebas de punción dactilar tomadas en el momento de la encuesta, y también se eliminaron los valores que eran muy discordantes. Las mediciones de glucosa se estratificaron aún más mediante muestras de glucosa aleatorias o en ayunas basadas en las respuestas del cuestionario de los participantes.

La presión arterial se seleccionó manualmente para individuos cuyos valores diferían en más de 20 unidades para las dos lecturas tomadas, para quienes la presión diastólica era mayor que la sistólica o para quienes los valores eran inusualmente altos o bajos (<30 o >300). En estos casos, ambas lecturas se verificaron manualmente y se descartaron las lecturas discordantes. Estos valores actualizados se fusionaron con las muestras restantes. También se generó un conjunto de fenotipos de presión arterial ajustados, ajustando para el tratamiento de la hipertensión. En aquellos individuos que se informó que estaban recibiendo algún tipo de tratamiento para la hipertensión, se añadieron 15 unidades a la presión arterial sistólica y 10 a la presión arterial diastólica (PAS_adj y PAD_adj)^{53,54}.

Los rasgos cuantitativos se normalizaron utilizando una transformada normal inversa antes de los análisis de rasgos complejos.

Para cada individuo, tenemos acceso a datos de diversos factores socioculturales, como el acceso a la atención médica y al agua potable, los ingresos anuales, el nivel educativo, si hablan una lengua indígena o no, y si viven en un entorno rural o urbano.

A las localidades se les asignaron valores de latitud, longitud y altitud (metros) utilizando datos del Instituto Nacional de Estadística y Geografía (INEGI) en México.

Selección de muestras y genotipado para el Proyecto MXB

Para seleccionar el subconjunto de muestras del biobanco a genotipar, primero identificamos el número total de localidades representadas en la colección de ADN extraídos (es decir, 898 sitios de reclutamiento). Luego asignamos una muestra a cada localidad en rondas aditivas consecutivas dirigidas a un tamaño de muestra promedio de 5 a 10 individuos, independientemente de la densidad de población. Las rondas iniciales se enriquecieron para las personas que informaron hablar una lengua indígena, y luego se incluyeron muestras

seleccionadas al azar hasta saturar la capacidad presupuestaria. Esta estrategia aseguró la maximización tanto de la cobertura geográfica como de la representación de las ancestrías indígenas, lo que resultó en un total de 6.144 muestras distribuidas en todo el país. Un subconjunto adicional de 87 muestras falló en el control de calidad del ADN o la hibridación durante la genotipificación, para un total de 6,057 muestras genotipadas con éxito. Las muestras se genotipificaron en el arreglo global multiétnica (MEGA, por sus siglas en inglés) de Illumina. El diseño de este arreglo fue dirigido previamente por C.R.G. y G.L.W. Varias propiedades colocan al arreglo MEGA como la opción ideal para la genotipificación de biobancos. Captura 1.748.250 SNP derivados de estudios de población mixta, por lo que es ampliamente aplicable en diversas poblaciones. El arreglo ha aumentado la cobertura de SNP en los loci MHC y KIR, un conjunto de marcadores de más de 30 000 SNP para la estimación de la ancestría, e incluye más de 17 000 variantes genéticas médica mente relevantes de GWAS y estudios clínicos anteriores. Tal amplitud de cobertura de la diversidad genómica proporciona un recurso cuantitativo integral de la variabilidad genética en este cohort.

Generación y control de calidad de datos genéticos MXB

Genome Studio se utilizó para convertir archivos de imagen RAW en archivos plink con información de genotipo RAW. Todos los SNP se orientaron a la cadena positiva y se eliminaron los SNP duplicados. Para los sitios con número de cromosoma faltante, posición física o ambos, actualizamos el mapa utilizando la información del nombre del SNP o mapeando su rsID utilizando dbSNP Build 151.

Eliminamos todos los individuos con >5% de datos de genotipo faltantes y todos los genotipos con >5% de individuos faltantes. Restringimos los análisis a autosomas y eliminamos todos los SNP monomórficos. Restringimos el análisis a los SNP bialélicos y eliminamos todos los SNP con una cadena ambigua para todos los análisis posteriores. Todos los individuos relacionados se detectaron usando plink (--Z-genome --min 0.5) después de filtrarlos por desequilibrio de ligamiento (--indep-pairwise 50 5 0.5). Se escribió una secuencia de comandos para encontrar y eliminar iterativamente a las personas relacionadas para obtener el conjunto de datos final de calidad controlada.

Fuentes y control de calidad para paneles de referencia

Se utilizaron paneles genéticos de referencia para diversos análisis de la estructura de la población. Utilizamos poblaciones globales del Proyecto 1000 Genomas (1KGP)⁴⁰ y el Proyecto de Diversidad del Genoma Humano (HGDP)⁵⁵, individuos zapotecos de Oaxaca de la Arquitectura de la Población utilizando el Estudio de Genómica y Epidemiología (PAGE)⁵⁶ e individuos indígenas de todo México del Proyecto de Diversidad de los Nativos Mexicanos (NMDP)¹² para los análisis de la estructura y ancestría de la población.

Para cada panel de referencia, restringimos el análisis a autosomas, eliminamos todos los SNP monomórficos, cambiamos todos los SNP a la cadena positiva y eliminamos los SNP con cadena ambigua.

Clasificación antropológica

Utilizamos un contexto antropológico y arqueológico para delinear diferentes regiones mesoamericanas¹⁰. Se utilizó la localidad de un individuo para ubicarlos en una de las siete regiones: el norte de México, el norte de Mesoamérica, el centro, occidente y Golfo de México, Oaxaca y la región maya en el sureste¹⁰. Esta clasificación se utilizó para visualizar y regionalizar algunos de los análisis de estructura e historia de la población, especialmente los relacionados con la subestructura genética indígena dentro de México.

Nota sobre ancestrías genéticas

La ancestría genética surge de un conjunto de caminos a través del gráfico de recombinación ancestral⁵⁷. En este estudio, obtenemos proxies para ancestrías genéticas utilizando ADMIXTURE²⁰ (ver a continuación). Como tal, estamos discretizando una cantidad continua con el fin de comprender los efectos de las diferentes historias demográficas sobre la variación genética y de rasgos complejos en MXB. El etiquetado y el uso de tales proxies de ancestría discretizados sigue siendo un tema polémico⁵⁸. Para aclarar el punto de que tales proxies son entidades especializadas en el mundo real, sino variables que utilizamos para los fines que acabamos de describir, optamos por referirnos a nuestros proxies ancestrales como de la región con cuyos individuos actuales se agrupan tales proxies. Por lo tanto, usamos, "ancestrías de las Américas", "ancestrías de Europa Occidental", "ancestrías de África Occidental", "ancestrías de Asia del Sur" y "ancestrías de Asia Oriental" en el texto, y versiones más cortas de los mismos para algunas figuras (A(Américas) y así sucesivamente).

Dichas regiones son útiles para nuestros análisis solo en la medida en que reflejan historias demográficas y ambientales que pueden afectar la variación genética y de rasgos complejos que nos interesan. Esta es solo una escala arbitraria para discretizar, y también consideramos los orígenes y las implicaciones de las variaciones ancestrales dentro de tales agrupaciones regionales en varios análisis, en los que llevamos a cabo la reducción de la dimensionalidad dentro de tales agrupaciones regionales (por ejemplo, MDS1 (A(Américas)) y MDS2(A(Américas))).

Aunque no se pretende, las agrupaciones utilizadas pueden parecer similares a las categorías raciales que se crearon en los últimos 500 años y se utilizaron para justificar la superioridad europea y la colonización de regiones globales, incluido el México actual^{58,59,60}. En México, tales categorías tienen una historia similar de racismo y eugenésia como en otras partes del mundo⁶¹. Rechazamos las categorizaciones jerárquicas fijas de los humanos, así como su uso para justificar la superioridad de un grupo sobre otro. Utilizamos proxies de ancestría que se estiman a partir de la ADMIXTURE utilizando agrupaciones no supervisadas, así como ejes de ancestría que resultan de la reducción de la dimensionalidad dentro de estas ancestrías, capturando la variación entre grupos de las Américas, por ejemplo. A pesar de la confluencia de ancestrías genéticas de todo el mundo en el México actual, los ancestrías genéticos en los humanos son continuos en el tiempo y el espacio y deben considerarse solo en esa complejidad y en diferentes escalas.

Análisis de la estructura de la población

Para los análisis de la estructura de la población, fusionamos el conjunto de datos MXB con filtro de control de calidad y los paneles de referencia utilizando plink. Repetimos algunos de los pasos de control de calidad en el conjunto de datos fusionado, eliminando cualquier SNP monomórfico o duplicado. También eliminamos individuos con >5% de datos de genotipo faltantes y genotipos con >5% de individuos faltantes para obtener el conjunto de datos fusionados limpios.

Realizamos dos conjuntos de análisis de componentes principales (PCA) y análisis ADMIXTURE²⁰. Uno se llevó a cabo en el conjunto de datos fusionado que incluye MXB, zapotecos de la Arquitectura de la Población utilizando el Estudio de Genómica y Epidemiología, y poblaciones globales del Proyecto 1000 Genomas y el Proyecto de Diversidad del Genoma Humano (Fig. 1b, Figs 6, 11 y 12 Suplementaria y Tabla suplementaria 3), y el otro se llevó a cabo en el conjunto de datos fusionado que incluye solo MXB e individuos indígenas del México actual del NMDP (Figs. suplementarias 8–10 y 13). El análisis de FST se llevó a cabo en todos los individuos MXB, así como solo en individuos MXB con un 90% de aancestría de las Américas según lo estimado a partir del análisis del ADMIXTURE (Figs. 15–18 suplementarias).

smartpca de Eigenstrat⁶² se utilizó para llevar a cabo el PCA. Los componentes principales generados por smartpca (Figs. 6 y 7) se usaron para llevar a cabo el análisis de aproximación y proyección de colector uniforme (UMAP) (Fig. 1c y Figs. 19 y 20 suplementarias)⁶³. El análisis FST se llevó a cabo utilizando smartpca.

Dada la gran pérdida de SNP debido al desequilibrio de ligamiento en nuestros individuos mexicanos, optamos por no filtrar el desequilibrio de ligamiento para los análisis de la estructura de la población presentados en este estudio. Repetimos el análisis en un conjunto de SNP filtrados para el desequilibrio de ligamiento y obtuvimos resultados similares (datos no mostrados). A menos que se indique lo contrario, dada la naturaleza mezclada de los individuos mexicanos, no eliminamos los SNP debido a la desviación del equilibrio de Hardy–Weinberg en el MXB, ya que se espera que muchos SNP estén fuera del equilibrio de Hardy–Weinberg debido a la mezcla y la estructura de la población.

También calculamos y visualizamos la estructura de la población utilizando el método de la ref.⁵ ('análisis arquetípico') con individuos del conjunto de datos MXB de calidad controlada e individuos de los 1000 Genomas, el Proyecto de Diversidad del Genoma Humano y la Arquitectura de la Población utilizando el Estudio de Genómica y Epidemiología como nuestro panel de referencia (Figuras 21–23 suplementarias). También realizamos el análisis utilizando solo el conjunto de datos MXB filtrados por control de calidad. En ambos análisis, los resultados del PCA se generaron solo una vez y se utilizaron como entrada para calcular los arquetipos de K = 3 a 10. Al informar los resultados, nos referimos a los 'arquetipos' en el análisis como 'fuentes', dado que la palabra arquetipos tiene connotaciones de tipos puros que no son necesarios para que el modelo se aplique a los datos genéticos de la población.

Análisis de subestructura poblacional

Se realizaron análisis para obtener ejes de variación genética o ancestría entre un grupo continental. Tales análisis también ayudan a interpretar los orígenes específicos de una ancestría presente en el México actual. Estos análisis se llevaron a cabo utilizando rfmix² para estimar la ancestría local a lo largo del genoma y pcamask¹⁹ para llevar a cabo un PCA específico de ancestría para las ancestrías originarias de la actual África. Durante el curso de este estudio, se publicaron métodos nuevos y mejorados para estimar la ancestrías local a lo largo del genoma (GNOMIX)³ y para llevar a cabo PCA específicos de la ancestría⁶⁴(Multiple Array Ancestry Specific Multidimensional Scaling, MAAS-MDS, un MDS diseñado para analizar muestras de varios arreglos de genotipado diferentes simultáneamente), lo que nos permite utilizar estas herramientas para el análisis de la variación de la ancestría dentro de las Américas para el análisis de rasgos complejos.

MAAS-MDS de ancestrías de las Américas

Para los análisis MAAS-MDS⁶⁴, utilizamos GNOMIX³ para la inferencia de ancestría local utilizando su modo "*Best*" preestablecido y luego enmascaramos los segmentos no indígenas. Para la referencia europea, utilizamos las cohortes poblaciones ibéricas en España (IBS) y británicas de Inglaterra y Escocia (GBR) de 1KGP (198 muestras) 40, para ancestrías de África, la cohorte Yoruba en Ibadan, Nigeria (YRI) de 1KGP (108 muestras), y para ancestrías de las Américas, peruano en Lima, Perú (PEL) de 1KGP (solo aquellas muestras con >95% de ancestría de las Américas) y los 50 genomas de individuos indígenas en todo México generados como parte del Proyecto MXB (79 muestras) ⁶⁵. Para el PEL, utilizamos un análisis de agrupamiento no supervisado con ADMIXTURE (K = 3) junto con IBS e YRI del 1KGP para encontrar aquellas muestras de PEL con >95% de asignación a un grupo no compartido con IBS o YRI; es decir, con >95% de ancestría de América. Los 50 genomas adicionales de MXB se seleccionaron para tener altas ancestrías indígenas como se describió anteriormente⁶⁵. Los genomas de referencia se fusionaron con cada arreglo dando como resultado 856,352 SNP en el arreglo 1 y 967,338 en el arreglo 2. El arreglo 1 incluyó 10 grupos indígenas de NMDP genotipados con el arreglo Affymetrix 6.0: Tarahumara, Huichol, Purepecha, Nahua, Totonaca, Mazateca, Zapoteca del Norte (del distrito de Villa Alta, Sierra del Norte en el estado de Oaxaca), Triqui, Tzotzil y Maya (del estado de Quintana Roo). El arreglo 2 incluyó a los 6.051 individuos del proyecto MXB genotipados con MEGA. El MAAS-MDS se aplicó a los segmentos de ancestría indígena americana (es decir, enmascarando componentes intercontinentales de origen africano y europeo) en ambas arreglos 1 y 2. El análisis se realizó utilizando distancias genéticas promedio por pares y considerando solo individuos con >20% de ancestría indígena americana, para generar ejes de MDS específicos de ancestría para ancestrías de las Américas en el MXB (Fig. 24 suplementaria).

asPCA de ancestrías de África

Llevamos a cabo este análisis en todos los individuos en el MXB con $\geq 5\%$ de ancestría de África estimada a partir del análisis de Admixture. Esto dio como resultado 1.965 individuos con ancestría originaria de la actual África. En este conjunto de individuos, utilizamos poblaciones del Proyecto 1000 Genomas (CEU: residentes de Utah (CEPH) con ancestrías del norte y oeste de Europa e YRI: Yoruba en Ibadan, Nigeria)⁴⁰ y el Estudio de Arquitectura de Poblaciones utilizando Genómica y Epidemiología (Zapotecas de Oaxaca)⁵⁶ para estimar la ancestría local utilizando rfmix². La región MHC se excluyó del análisis. Los SNP fuera del equilibrio de Hardy–Weinberg se eliminaron de cada uno de los paneles de referencia (10^{-3}) y el subconjunto MXBAFR (10^{-8}) de antemano. Este conjunto de datos se fusionó con un panel de referencia subcontinental que cubre una variedad de grupos en el África actual²¹. Pcamask¹⁹ se utilizó para enmascarar todas las ancestrías distintas de las africanas y para generar componentes principales específicos de ancestrías africanas en el MXB (Fig. 14 suplementaria).

Análisis de historias poblacionales

Estimación específica de ancestría de las trayectorias efectivas del tamaño de la población

Se llevaron a cabo análisis de la historia de la población utilizando un enfoque que utiliza segmentos de identidad por descendencia específicos de ancestría (IBD) en todo el conjunto de datos de MXB y en individuos pertenecientes a cada una de las regiones mesoamericanas (Fig. 2 y Figs. 25–29 suplementarias). Los segmentos de IBD del genoma se pueden utilizar para estimar el tamaño efectivo de la población (N_e) durante miles de años en el pasado³⁰. Estos segmentos de IBD se pueden superponer aún más con los tracts de ancestría local para obtener tracts de IBD específicos de la ancestría para estimar el tamaño de la población de una manera específica de la ancestría para una cohorte mixta (este enfoque se ha denominado como IBDNe)⁴.

Para este análisis, el MXB se fusionó con poblaciones del Proyecto 1000 Genomas (CEU: residentes de Utah (CEPH) con ancestría del norte y oeste de Europa e YRI: Yoruba en Ibadan, Nigeria)⁴⁰ y el Estudio de Arquitectura de Poblaciones utilizando Genómica y Epidemiología (Zapotecas de Oaxaca)⁵⁶. Los SNP en cada población se filtraron previamente para el equilibrio de Hardy–Weinberg (10^{-5} para los grupos de referencia y 10^{-10} para las muestras de MXB). La región MHC se excluyó del análisis. Repetimos algunos de los pasos de control de calidad en el conjunto de datos fusionado, eliminando cualquier SNP monomórfico o duplicado. También eliminamos individuos con $>5\%$ de datos de genotipo faltantes y genotipos con $>5\%$ de individuos faltantes para obtener el conjunto de datos fusionados limpios.

Seguimos un proceso computacional recomendado por los desarrolladores de asIBDNe para llamar a los segmentos de IBD y la ancestría local a lo largo del genoma. Utilizamos beagle (beagle.25Nov19.28d.jar)⁶⁶ para poner en fase los datos, refined-ibd

(refined-ibd.17Jan20.102.jar)⁶⁷ para llamar a IBD y merge-ibd-segments (merge-ibd-segments.17Jan20.102.jar) para eliminar roturas y segmentos cortos en los segmentos de IBD, eliminando las brechas entre los segmentos de IBD que tienen como máximo un homocigoto discordante y que tienen menos de 0.6 cM de longitud. La ancestría local se estimó utilizando rfmix. La salida de rfmix se reajustó para que coincidiera con la fase original. asIBDNe (ibdne.19Sep19.268.jar) se ejecutó para estimar los tamaños de población específicos de la ancestría utilizando un umbral de longitud de IBD de 2 cM.

AdmixtureBayes

En este estudio, utilizamos AdmixtureBayes⁶ para generar, analizar y trazar gráficos de mezcla para una muestra de 6.011 individuos del MXB (Datos extendidos Fig. 1a y Fig. 30 suplementaria). Nuestro enfoque fue inferir la historia demográfica de los grupos indígenas en México, por lo que utilizamos solo las frecuencias alélicas de las porciones indígenas de los genomas MXB. En particular, utilizamos GNOMIX para la inferencia de ancestría local como se describe en la sección titulada "MAAS-MDS sobre ascendencias de las Américas" en los Métodos, y enmascaramos los segmentos no indígenas.

Agrupamos a los individuos sobre la base de las regiones mesoamericanas de México, para comprender la variación de las historias demográficas indígenas en todo el país. Utilizamos la etnia china Han como un grupo externo para las ascendencias indígenas.

Usando AdmixtureBayes, inferimos los eventos de división y los eventos de mezcla que se han producido en el MXB. Utilizamos los parámetros predeterminados para generar el gráfico de mezcla con la excepción del número de cadenas e iteraciones, que establecimos en un valor más alto de 16 (--MCMC_Chains 16) y 20.000 (--n 20000) para garantizar la convergencia; también utilizamos el indicador -slower, que permite el cálculo de la información necesaria para trazar los árboles superiores, y un *burn-in period* correspondiente a la mitad de las muestras. Trazamos el árbol con las probabilidades posteriores más altas, lo que proporciona una representación visual de los eventos de mezcla inferidos y nos permite explorar la incertidumbre en las inferencias. Se pueden encontrar más detalles sobre el método AdmixtureBayes y el uso anterior en el documento correspondiente⁶.

ROH

El conjunto de datos MXB se filtró para el desequilibrio de ligamiento utilizando plink (--indep-pairwise 50 5 0.9). Los ROH se estimaron utilizando Plink (--homozyg) identificando 349,400 ROH. Estimamos el número de ROH transportados por un individuo (nROH) y la suma total de ROH en un individuo en kilobases (sROH o sumROH) (Fig. 3 y Figs. 31–33 suplementarias). Los ROH se dividieron en pequeños, medianos y grandes de acuerdo con el marco teórico en la ref.³⁷. Se utilizaron códigos de Python para categorizar ROH por longitud y para superponer ROH con llamadas de ancestría local de rfmix para obtener estadísticas de resumen de ROH específicas por ancestría (Tabla suplementaria 5). Las llamadas de ancestría local fueron las mismas que las utilizadas para el análisis de

asIBDNe. Un total de 38,340 ROH no se superpusieron a una asignación de ancestría local homocigótica y se eliminaron de este análisis; los 311,060 restantes que se superpusieron a una asignación de ancestría local homocigótica se mantuvieron. También utilizamos un código de Python para calcular el número de ROH en los puntos de cambio de ancestría(58 ROH o 0,00019 de todos los ROH cayeron dentro de un cambio de ancestría y también se excluyeron del análisis).

Los ROH también se correlacionaron con el año de nacimiento en el MXB (Fig. 3b) y se utilizaron como variable en el análisis de modelo mixto de rasgos complejos. Para el análisis del año de nacimiento, eliminamos las dos primeras décadas, ya que cada año tiene menos de 15 individuos muestreados en este período. El año de nacimiento también se correlacionó directamente con las ancestrías de las Américas (inferidas mediante ADMIXTURE) en localidades rurales y urbanas por separado. Los ROH también se correlacionaron con las ancestrías globales por individuo estimadas a partir del análisis de mezcla (Fig. 3a y Fig. 3 suplementaria). Se utilizó una secuencia de comandos R para analizar las distribuciones de la suma de ROH por geografía (Figs. 31 y 33 suplementarias).

Análisis de carga mutacional

Las variantes se anotaron en función de si eran ancestrales o derivadas, y su efecto funcional en función de su localización en un gen o genoma. Los alelos ancestrales para cada SNP en el MXB se dedujeron utilizando la tubería EPO del Proyecto 1000 Genomas. Se utilizó Variant Effect Predictor⁶⁸ para anotar el efecto de una variante utilizando la base de datos humdiv y eligiendo una consecuencia (o transcripción) por variante de acuerdo con un criterio que incluye el estado canónico de la transcripción, la anotación de la isoforma APPRIS, el nivel de soporte de la transcripción, el biotipo de transcripción (se prefiere ‘protein_coding’) y el rango de consecuencias que prefiere un alto impacto.

La carga mutacional se define como la suma de alelos derivados portados por un individuo. Se utilizó un tubería computacional utilizando vcftools, python, linux y R para calcular la carga mutacional en diferentes clases de variantes y en diferentes umbrales de frecuencia de alelos derivados. Calculamos una carga mutacional rara (frecuencia alélica derivada $\leq 5\%$) o una carga mutacional general considerando todas las frecuencias alélicas. Nuestro pipeline utilizó los paquetes R matrixStats, dplyr y ggplot2. Correlacionamos la carga mutacional con el porcentaje de ancestría global de diferentes orígenes continentales actuales en todos los individuos. Las estimaciones de ancestría fueron del análisis de mezcla. Calculamos la correlación de Spearman y el valor P (Fig. 4).

Este análisis se repitió en la cohorte 1000 Genomes Project Mexican Ancestry en Los Ángeles, California (MXL) (Fig. 39 suplementaria). Esto fue para verificar si el efecto que estábamos observando se debía al sesgo de determinación en el arreglo MEGAex que cubre menos variantes raras predominantemente nativas del área que es México hoy en día. Las secuencias de todo el genoma del Proyecto 1000 Genomas nos permitieron descartar esto.

Las estimaciones de ancestría se generaron utilizando Un ADMIXTURE con paneles de referencia de 1000 genomas (CEU: residentes de Utah (CEPH) con ancestría del norte y oeste de Europa, GBR: británicos en Inglaterra y Escocia, YRI: yoruba en Ibadan, Nigeria y PEL: peruano en Lima, Perú) y 50 secuencias de genoma completo de individuos indígenas en todo México generadas como parte del Proyecto MXB⁶⁵. Se utilizó el predictor de efecto variante para anotar los SNP, y la carga mutacional se calculó de la misma manera. La categoría deletéreas incluye los siguientes términos de consecuencia: variante aceptora de empalme, variante donante de empalme, ganancia de codón de paro, pérdida de codón de paro, pérdida de codón de inicio.

Análisis GWAS

Definiciones de fenotipo y control de calidad

Los fenotipos binarios relacionados con la salud se definieron sobre la base de las respuestas al cuestionario. Los casos se definieron sobre la base de una respuesta positiva a las preguntas del cuestionario. Los controles fueron aquellos que respondieron con un "no". Se excluyeron las personas que respondieron con "no lo sé", "prefiero no responder" o "sin respuesta" (Tabla suplementaria 6). Además, los casos de artritis se definieron como cualquier individuo con artritis gotosa, artritis reumatoide y/u otras formas de artritis. Se definieron dos fenotipos de hipertensión: Hipertensión_1, basada en un diagnóstico de hipertensión; e Hipertensión_2, que además tuvo en cuenta las lecturas de presión arterial. Los casos se definieron sobre la base de un diagnóstico de hipertensión, medicación o lecturas de presión arterial superiores a 140/90.

Los rasgos cuantitativos se midieron como se describió anteriormente⁵¹. Los datos se filtraron para eliminar valores atípicos con errores aparentes en la entrada de datos y valores negativos (<0) en función de la densidad de distribución en el conjunto de datos. La altura se limitó a los participantes con medidas entre 100 y 200 cm; el peso se limitó a entre 25 y 300 kg. Los niveles de glucosa y glucosa en ayunas se compararon con las pruebas de punción dactilar tomadas en el momento de la encuesta y se eliminaron los valores que eran muy discordantes. Las mediciones de glucosa en ayunas se definieron en función de si los participantes habían comido en las 8–12 h previas a la toma de muestras.

La presión arterial se seleccionó manualmente para individuos cuyos valores diferían en más de 20 unidades para las dos lecturas tomadas, para quienes la presión diastólica era mayor que la sistólica o para quienes los valores eran inusualmente altos o bajos (<30 o >300). En estos casos, ambas lecturas se verificaron manualmente y se descartaron las lecturas discordantes. Estos valores actualizados se fusionaron con las muestras restantes. Para GWAS, se utilizó el primer conjunto de lecturas a menos que se eliminara durante el proceso de control de calidad, en cuyo caso se utilizó el segundo conjunto de lecturas, si estaba disponible. También se generó un conjunto de fenotipos de presión arterial ajustados, ajustando para el tratamiento de la hipertensión. En aquellos individuos que se informó que estaban recibiendo

algún tipo de tratamiento para la hipertensión, se añadieron 15 unidades a la presión arterial sistólica y 10 a la presión arterial diastólica.

Análisis de Asociación del Genoma Completo (GWAS)

Los análisis de GWAS para los rasgos binarios y cuantitativos se llevaron a cabo con regenie (v3.1.3)⁶⁹. Antes de GWAS, se eliminaron las personas con sexo no coincidente o IBD > 0,9. Los rasgos cuantitativos se normalizaron inversamente antes del análisis. Solo los rasgos de casos y controles con más de 100 casos se tomaron para su análisis. Para todos los análisis, se incluyeron como variables la edad, el sexo y los cuatro primeros componentes principales. Para el colesterol, triglicéridos, HDL, LDL, hipertensión y glucosa en ayunas, el IMC también se incluyó como variable.

Puntuación poligénica GWAS

GWAS se llevó a cabo en un subconjunto aleatorio de 4.000 individuos con datos de genotipo disponibles, como se describió anteriormente. Para los rasgos cuantitativos, los valores brutos se normalizaron nuevamente dentro del subconjunto seleccionado antes del análisis.

Mapeo fino de loci significativos de GWAS

Los SNP de asociación principal y los grupos causales potenciales se definieron utilizando FINEMAP (v1.3.1; R² = 0.7; factor de Bayes ≥ 2) de SNP dentro de cada una de estas regiones sobre la base de estadísticas de resumen para cada uno de los rasgos asociados⁷⁰. A continuación, se utilizó FUMA SNP2GENE para identificar los genes más cercanos a cada locus sobre la base del desequilibrio de ligamiento calculado utilizando las poblaciones EUR de 1000 genomas, y explorar las asociaciones previamente informadas en el catálogo GWAS^{40,71} (Tabla suplementaria 7).

Análisis de puntuación poligénica

Calculamos las puntuaciones poligénicas utilizando estadísticas de plink y resumen del MXB GWAS realizado en 4.000 individuos como se describió anteriormente⁷². Calculamos las puntuaciones de las 1.778 personas restantes. También calculamos las puntuaciones para los mismos individuos utilizando las estadísticas resumidas de UKB GWAS de panastría (<https://pan.ukbb.broadinstitute.org>)^{7,8} (Fig. 41 suplementaria). El desequilibrio de ligamiento se explicó por la aglutinación usando plink usando un valor r² de 0.1, y las puntuaciones poligénicas se calcularon usando SNP significativos en cinco umbrales de valor P diferentes (0.1, 0.01, 0.001, 0.00001 y 10-8) con el modificador de suma de puntuación - (que da la suma de todos los alelos asociados en un umbral de valor P ponderado por sus tamaños de efecto estimados). Probamos el rendimiento de predicción de las puntuaciones poligénicas calculando la correlación de Pearson entre el valor del rasgo y la puntuación poligénica

(Tablas suplementarias 8 y 9). Además, creamos un modelo lineal nulo para cada rasgo que incluye la edad, el sexo y diez componentes principales como variables. Creamos un segundo modelo de puntuación poligénica sumando la puntuación poligénica al modelo nulo. Calculamos la r^2 de la puntuación poligénica tomando la diferencia entre la r^2 del modelo de puntuación poligénica y la r^2 del modelo nulo. En general, la predicción basada en MXB se mejora mediante el uso de todos los SNP asociados a $P < 0.1$ y el uso de datos imputados por TOPMed, mientras que la predicción basada en UKB muestra su mejor rendimiento utilizando solo SNP significativos en todo el genoma ($a 10^{-8}$ o 10^{-5}) y solo datos genotipados (Datos extendidos Fig. 1b y Tablas suplementarias 8 y 9).

Modelos complejos de variación de rasgos

Para evaluar los factores involucrados en la creación de la variación de rasgos complejos, llevamos a cabo un análisis de modelos mixtos utilizando el paquete lme4qtl R para todos los rasgos cuantitativos. lme4qtl permite la creación de modelos flexibles con múltiples efectos aleatorios⁷³.

Consideramos varias variables genéticas y ambientales como predictores fijos de la variación de rasgos complejos. Las variables genéticas incluyeron puntuaciones poligénicas calculadas utilizando estadísticas de resumen de UKB (SNP significativos a $P < 10^{-8}$) para cada rasgo, ancestrías genéticas estimadas a partir del ADMIXTURE, ejes continuos de variación de ancestrías estimados utilizando MAAS-MDS y ROH (cantidad de ROH transportada en un genoma individual en kilobases). También consideramos variables biogeográficas como latitud, longitud y altitud (metros). Consideramos variables demográficas de edad y sexo. Por último, consideramos variables socioculturales: logro educativo (que muestra una correlación positiva con los niveles de ingresos (Fig. 52 suplementaria); sin embargo, los niveles de ingresos están disponibles solo para un tercio de los individuos); si hablan una lengua indígena o no como indicador de la experiencia diferencial de discriminación y cultura; y si viven en un entorno urbano o rural. El IMC se incluyó como variable para todos los rasgos cuantitativos excepto la altura, el IMC y la creatinina (Fig. 5 y Figs. 53–58 suplementarias). Para facilitar la interpretación de los coeficientes del modelo mixto para los predictores numéricos y binarios considerados conjuntamente, estandarizamos las variables predictoras de la siguiente manera⁷⁴. Para hacer que los coeficientes de los predictores numéricos sean comparables a los de los predictores binarios no transformados, dividimos cada variable numérica por dos veces su desviación estándar⁷⁴. Centramos tanto los predictores binarios como los numéricos. Todas las variables mencionadas anteriormente son significativas cuando se modelan conjuntamente para al menos un rasgo probado, lo que justifica su uso en el modelo completo.

También incluimos dos predictores aleatorios en nuestro modelo. Estos son: la estructura de varianza definida por la matriz de relación genética; y la localidad de donde proviene el individuo para capturar cualquier otra variación ambiental (como la dieta) no capturada por los predictores fijos.

La matriz de relación genética se generó utilizando el paquete GENESIS R utilizando coeficientes de parentesco. Como las estimaciones de parentesco se pueden inflar bajo la presencia de la estructura y mezcla de la población, obtuvimos coeficientes de parentesco para la matriz de relación genética de la siguiente manera: (1) se utilizó PC-air⁷⁵ para obtener componentes principales que capturan la ancestría y no el parentesco (este procedimiento utilizó coeficientes de parentesco estimados utilizando KING⁷⁶ como entrada para dividir muestras en un conjunto relacionado (5,562) y no relacionado (271) (utilizando el umbral de parentesco 0.044) y llevando a cabo PCA en el conjunto no relacionado); (2) se utilizó PC-relate⁷⁷ para obtener coeficientes de parentesco que capturan la relación pero no la ancestría (este método utiliza los componentes principales representativos de la ancestría de (1) para corregir la estructura de la población antes de calcular los coeficientes de parentesco).

Para este análisis, eliminamos variantes raras ($MAF < 5\%$), regiones con desequilibrio de ligamiento de largo alcance conocido^{78,79} y variantes en alto desequilibrio de ligamiento ($r^2 > 0.1$ en una ventana de 50 kb y una ventana deslizante de 1 variante).

Para tener en cuenta las pruebas de significancia múltiple, la tasa de descubrimiento falso se controló a 0.05 utilizando el enfoque de Benjamini–Hochberg⁵⁰.

Las frecuencias de la variante ABCA1 se calcularon utilizando plink en individuos del MXB estratificados por proxies de ancestría del ADMIXTURE o por los niveles de colesterol HDL (Fig. 59 suplementaria).

Los mapas de México para visualizar las distribuciones de rasgos se crearon utilizando el paquete mxmaps R (Fig. 43 suplementaria). Se utilizó Variog del paquete GeoR R para calcular variogramas sobre rasgos complejos, con longitud y latitud utilizadas para calcular la distancia (Fig. 51 suplementaria).

Inclusión y ética

Las muestras se recolectaron como parte de la Encuesta Nacional de Salud 2000 (ENSA 2000) realizada por el Instituto Nacional de Salud Pública (INSP) en todo México. La ENSA 2000 se llevó a cabo siguiendo los principios éticos más estrictos y de acuerdo con la Declaración de Helsinki de Estudios Humanos. Se obtuvo el consentimiento informado de todos los participantes después de una amplia participación de la comunidad. Las Encuestas Nacionales de Salud se han realizado periódicamente en México desde 1988, por lo que la comunidad participa en el estudio y es receptiva a las visitas domiciliarias del personal del INSP y los equipos de trabajo de campo. Como se describe en la metodología original⁵¹, la ENSA 2000 implicó una visita de 2 horas a cada hogar. Antes de la contratación, el equipo se reunió con los líderes políticos, religiosos y comunitarios de cada localidad para comunicar la naturaleza del estudio, responder a todas las preguntas e interactuar con la comunidad. Este proceso de participación comunitaria fue esencial en todos los sitios de reclutamiento, con

énfasis en las comunidades indígenas y rurales para garantizar la comprensión del estudio. Los ADN extraídos se han almacenado y mantenido en el INSP (Cuernavaca, México), y las muestras seleccionadas se genotiparon en la Unidad de Genómica Avanzada del CINVESTAV (Irapuato, México) a través de un acuerdo de colaboración. Los datos han sido analizados conjuntamente, promoviendo el liderazgo local y la participación de investigadores y aprendices mexicanos. El proyecto fue revisado y aprobado por el Comité de Ética en Investigación y el Comité de Bioseguridad del INSP (Aprobaciones de la Junta de Revisión Institucional CI: 1479 y CB: 1470). Para el presente proyecto, los datos de identificación personal se eliminaron del conjunto de datos.

Resumen de los informes de investigación

Más información sobre el diseño de la investigación está disponible en el Resumen de Informes de Investigación de la Naturaleza vinculado a este artículo.

Disponibilidad de los datos

Los conjuntos de datos de genotipo y fenotipo para los 6.057 individuos recién genotipados del Proyecto Biobanco MX están disponibles en el Archivo Europeo de Genoma-fenoma (EGA) a través de un Acuerdo de Acceso a Datos con el Comité de Acceso a Datos (número de acceso de EGA para el estudio: EGAS00001005797; conjuntos de datos: EGAD00010002361 (Mexican_Biobank_Genotypes) y EGAD00001008354 (Mexican Biobank 50 Genomes)). Se puede acceder a los datos solo para investigación académica y uso no comercial. Las estadísticas resumidas de GWAS generadas como parte de este estudio están disponibles en <https://doi.org/10.5281/zenodo.7420254>.

Referencias

1. Mills, M. C. & Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat. Genet.* 52, 242–243 (2020).
2. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* 93, 278–288 (2013).
3. Hilmarsdóttir, H., Kumar, A. S., Rastogi, R. & Bustamante, C. D. High resolution ancestry deconvolution for next generation genomic data. Preprint at bioRxiv <https://doi.org/10.1101/2021.09.19.460980> (2021).
4. Browning, S. R. et al. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* 14, e1007385 (2018).
5. Gimbernat-Mayol, J., Mantes, A. D., Bustamante, C. D., Montserrat, D. M. & Ioannidis, A. G. Archetypal analysis for population genetics. *PLoS Comput. Biol.* 18, e1010301 (2022).
6. Nielsen, S. V. et al. Bayesian inference of admixture graphs on Native American and Arctic populations. *PLoS Genet.* 19, e1010410 (2023).
7. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209 (2018).
8. Pan-Ancestry Genetic Analysis of the UK Biobank <https://pan.ukbb.broadinstitute.org/> (Pan-UK Biobank, accessed date 2 October 2022).
9. Coe, M. D., Urcid, J. & Koontz, R. Mexico: from the Olmecs to the Aztecs (Thames & Hudson, 2013).
10. Vela, E. Áreas culturales: Oasisamérica, Aridamérica y Mesoamérica. *Arqueol. Mex* 82, 28–29 (2018).
11. Mendoza, R. G. in *The Oxford Encyclopedia of Mesoamerican Culture* Vol. 2 (ed. Carrasco, D.) 222–226 (2001).
12. Moreno-Estrada, A. et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344, 1280–1285 (2014).
13. García-Ortiz, H. et al. The genomic landscape of Mexican Indigenous populations brings insights into the peopling of the Americas. *Nat. Commun.* 12, 5942 (2021).

14. Romero-Hidalgo, S. et al. Demographic history and biologically relevant genetic variation of Native Mexicans inferred from whole-genome sequencing. *Nat. Commun.* 8, 1005 (2017).
15. Ávila-Arcos, M. C. et al. Population history and gene divergence in native Mexicans inferred from 76 human exomes. *Mol. Biol. Evol.* 37, 994–1006 (2020).
16. Rodríguez-Rodríguez, J. E. et al. The genetic legacy of the Manila galleon trade in Mexico. *Phil. Trans. R. Soc. B* 377, 20200419 (2022).
17. Spear, M. L. et al. Recent shifts in the genomic ancestry of Mexican Americans may alter the genetic architecture of biomedical traits. *Elife* 9, e56029 (2020).
18. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* 538, 161–164 (2016).
19. Moreno-Estrada, A. et al. Reconstructing the population genetic history of the Caribbean. *PLoS Genet.* 9, e1003925 (2013).
20. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664 (2009).
21. Patin, E. et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356, 543–546 (2017).
22. Trans-Atlantic Slave Trade Database <https://www.slavevoyages.org/> (Slave Voyages, accessed date 15 November 2021).
23. Seijas, T. Asian Slaves in Colonial Mexico: from Chinos to Indians (ed. Klein, H. S.) (Cambridge Univ. Press, 2014).
24. Chávez, C. P. M. El alcalde de los chinos en la Provincia de Colima durante el siglo XVII: un sistema de representación en torno a un oficio. *Let. Hist.* 1, 95–115 (2009).
25. Keresey, D. O. La esclavitud Asiática en el virreinato de la Nueva España, 1565–1673. *Hist. Mex.* 61, 5–57 (2011).
26. Carrillo, R. Asia llega a América. Migración e influencia cultural asiática en Nueva España (1565–1815). *Asiadémica* 3, 81–98 (2014).
27. Mishima, M. E. O. Siete Migraciones Japonesas en México: 1890–1978 (El Colegio de Mexico, 1982).
28. Augustine-Adams, K. Prohibir el mestizaje con chinos: solicitudes de amparo, Sonora, 1921–1935. *Rev. Indias* 72, 409–432 (2012).
29. Guillén, M. L. Vivir para trabajar. La inserción laboral de los inmigrantes chinos en Chiapas, siglos XIX y XX. *Studium: Revista Humanidades* 19, 113–140 (2013).
30. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* 97, 404–418 (2015).
31. Wang, R. J., Al-Saffar, S. I., Rogers, J. & Hahn, M. W. Human generation times across the past 250,000 years. *Sci Adv.* 9, eabm7047 (2023).
32. Gugliotta, G. The Maya: glory and ruin. *The National Geographic Magazine* 212, 68–109 (August 2007).
33. Diehl, R. A. *The Olmecs: America's First Civilization* (Thames & Hudson, 2004).
34. Marcus, J. & Flannery, K. in *The Cambridge History of the Native Peoples of the Americas* (eds Adams, R. E. W. & MacLeod, M. J.) 358–406 (Cambridge Univ. Press, 2000).
35. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* 19, 220–234 (2018).
36. Clark, D. W. et al. Associations of autozygosity with a broad range of human phenotypes. *Nat. Commun.* 10, 4957 (2019).
37. Ringbauer, H., Novembre, J. & Steinrück, M. Parental relatedness through time revealed by runs of homozygosity in ancient DNA. *Nat. Commun.* 12, 5425 (2021).
38. Henn, B. M. et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl Acad. Sci. USA* 113, E440–E449 (2015).
39. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating mutation load in human genomes. *Nat. Rev. Genet.* 16, 333–343 (2015).
40. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
41. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* 46, 220–224 (2014).
42. Do, R. et al. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* 47, 126–131 (2015).
43. Acuña-Alonso, V. et al. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. *Hum. Mol. Genet.* 19, 2877–2885 (2010).
44. Robinson, M. R. et al. Evidence of directional and stabilizing selection in contemporary humans. *Proc. Natl Acad. Sci. USA* 115, E4732 (2018).
45. Simons, Y. B., Mostafavi, H., Smith, C. J., Pritchard, J. K. & Sella, G. Simple scaling laws control the genetic architectures of human complex traits. Preprint at bioRxiv <https://doi.org/10.1101/2022.10.04.509926> (2022).
46. Malawsky, D. S. et al. Influence of autozygosity on common disease risk across the phenotypic spectrum. Preprint at medRxiv <https://doi.org/10.1101/2023.02.01.23285346> (2023).
47. Barquera, S. & Rivera, J. A. Obesity in Mexico: rapid epidemiological transition and food industry interference in health policies. *Lancet Diabetes Endocrinol.* 8, 746–747 (2020).
48. Mendoza-Caamal, E. C. et al. Metabolic syndrome in indigenous communities in Mexico: a descriptive and cross-sectional study. *BMC Public Health* 20, 339 (2020).
49. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* 367, eaay5012 (2020).

50. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300 (1995).
51. Sepúlveda, J. et al. Diseño y metodología de la Encuesta Nacional de Salud 2000. *Salud Pública Méx.* 49, s427–s432 (2007).
52. Gamboa-Meléndez, M. A. et al. Contribution of common genetic variation to the risk of type 2 diabetes in the Mexican Mestizo population. *Diabetes* 61, 3314–3321 (2012).
53. Warren, H. R. et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* 49, 403–415 (2017).
54. Hoffmann, T. J. et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat. Genet.* 49, 54–64 (2017).
55. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104 (2008).
56. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518 (2019).
57. Mathieson, I. & Scally, A. What is ancestry? *PLoS Genet.* 16, e1008624 (2020).
58. Lewis, A. C. F. et al. Getting genetic ancestry right for science and society. *Science* 376, 250–252 (2022).
59. Saini, A. Superior: the Return of Race Science (Beacon, 2019).
60. Yudell, M. *Race Unmasked: Biology and Race in the Twentieth Century* (Columbia Univ. Press, 2014).
61. Suárez y López Guazo, L. L. *Eugenésia y Racismo en México* (UNAM, 2005).
62. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909 (2006).
63. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet.* 15, e1008432 (2019).
64. Ioannidis, A. G. et al. Native American gene flow into Polynesia predating Easter Island settlement. *Nature* 583, 572–577 (2020).
65. Jiménez-Kaufmann, A. et al. Imputation performance in Latin American populations: improving rare variants representation with the inclusion of Native American genomes. *Front. Genet.* 12, 719791 (2022).
66. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097 (2007).
67. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* 194, 459–471 (2013).
68. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122 (2016).
69. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* 53, 1097–1103 (2021).
70. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501 (2016).
71. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1826 (2017).
72. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, s13742-015-0047-8 (2015).
73. Ziyatdinov, A. et al. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics* 19, 68 (2018).
74. Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* 27, 2865–2873 (2008).
75. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39, 276–293 (2015).
76. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873 (2010).
77. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* 98, 127–148 (2016).
78. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* 83, 132–135 (2008).
79. Tang, H. et al. Response to Price et al. *Am. J. Hum. Genet.* 83, 135–139 (2008).