Análisis de la diversidad genómica en las poblaciones mestizas mexicanas para desarrollar medicina genómica en México

Irma Silva-Zolezzi[1], Alfredo Hidalgo Miranda[1], Jesús Estrada Gil[1], Juan Carlos Fernández López, Laura Uribe Figueroa, Alejandra Contreras, Eros Balam Ortiz, Laura del Bosque Plata, César Lara, David Velázquez Fernández, Rodrigo Goya, Enrique Hernández Lemus, Carlos Dávila, Eduardo Barrientos, Santiago March, Gerardo Jiménez-Sánchez[2]

Instituto Nacional de Medicina Genómica, Periférico Sur No. 4124, Torre Zafiro II, 6to. Piso, Col. Jardines del Pedregal, México D.F. 01900, México
Communicated by Eric S. Lander, The Broad Institute

México está desarrollando las bases para la medicina genómica con el fin de mejorar el cuidado de la salud de su población. El estudio amplio de la diversidad genética y las estructuras de desequilibrio de ligamiento de diferentes poblaciones, ha hecho posible el desarrollo de estrategias de captura o "tagging" de variabilidad genética, así como de imputación de genotipos, para evaluar integralmente la variabilidad común en estudios de asociación en enfermedades comunes. Nosotros evaluamos el beneficio de un mapa de haplotipos mexicano para la identificación de genes relacionados a enfermedades comunes en mexicanos. Con este objetivo se caracterizaron la diversidad genética, los patrones de LD y la proporción de haplotipos compartidos utilizando datos genómicos de mestizos mexicanos provenientes de regiones con diferente historia de mestizaje y distintas dinámicas poblacionales. La composición ancestral de los mestizos fue evaluada incluyendo datos de un grupo amerindio mexicano y del HapMap. Nuestros resultados apoyan la existencia de diferencias genéticas regionales en México que deben ser consideradas en el diseño y análisis de estudios de asociación. Adicionalmente, dan sustento al hecho de que un mapa de haplotipos de la población mestiza mexicana podría reducir el número de tag SNPs requerido para caracterizar la variabilidad genética común en esta población. Este estudio es uno de los primeros esfuerzos de genotipificación amplia del genoma en una población de reciente mestizaje en América Latina .

mestizaje | variación genética | genética de población | captura de SNPs

Más de 560 millones de personas viven en los países de Latinoamérica y la Oficina de Censos de los Estados Unidos, estimó en 2007 que la población latina en este país había alcanzado un total de 45.5 millones, posicionándola como el grupo minoritario más grande y de crecimiento más acelerado en los Estados Unidos. Al igual que otras poblaciones latinas, los mestizos mexicanos constituyen una población mestiza de reciente formación conformada principalmente por orígenes ancestrales amerindio y europeos, y en menor proporción africano. Aunque el tamaño y la diversidad de las poblaciones latinas plantean diversos y serios desafíos para los estudios genéticos (1), también representan un poderoso recurso para el análisis de las bases genéticas de enfermedades complejas (2). En los últimos cinco años, México se ha comprometido a desarrollar infraestructura humana y tecnológica para la genómica como parte del desarrollo de una plataforma nacional de medicina genómica para mejorar la atención de la salud de la población Mexicana (3-6). Esta situación en conjunto con el hecho de que México tiene una población de ~105 millones de habitantes que incluye más de 60 grupos amerindios y una historia compleja de mestizaje, convierten a México en un país ideal en dónde realizar análisis genómicos en enfermedades complejas.

En la actualidad dos abordajes para la identificación de genes asociados a enfermedades complejas han sido exitosos: son los estudios de asociación de genoma completo y el mapeo por mestizaje (GWAS y AM, por sus siglas en inglés). El primero se basa en una eficiente captura ("tagging") de polimorfismos de un sólo nucleótido (SNPs, por sus siglas en inglés) (7, 8), y el segundo en la disponibilidad de paneles de marcadores distribuidos a lo largo del genoma con diferencia de frecuencias entre poblaciones ancestrales (AIMs, por sus siglas en inglés) (9, 10). En poblaciones sin una representación completa en el HapMap (11), como es el caso de los Latinos, existen limitaciones para la eficiente evaluación de la diversidad genética mediante estrategias de captura (tagging), debido a la necesidad de emplear un mayor número de marcadores para poder obtener un poder estadístico relativo similar al que se logra en asiáticos y europeos (12), y al desconocimiento acerca de los patrones de desequilibrio de ligamiento (LD, por sus siglas en inglés) específicos de estas poblaciones (13). Adicionalmente, en los GWAS los resultados falsos positivos por efecto de la estructura poblacional son minimizados al excluir a individuos con diferencias ancestrales (7). Lo anterior no es práctico en estudios en latinos, como los mexicanos, si se considera que el 80% de la población está constituida por mestizos, los cuales presentan notables diferencias individuales en constituciones ancestrales (2). A pesar de que recientemente se han desarrollado algunos paneles de SNPs para hacer AM en poblaciones latinas (14-16), la información detallada acerca de la diversidad genómica de poblaciones mestizas e indígenas aún es limitada (17, 18). Estudios recientes en poblaciones latinoamericanas han demostrado la existencia de patrones de contribución ancestral diferenciales entre e intra grupos, que correlacionan con la densidad poblacional indígena antes de la conquista de América y con los patrones de crecimiento demográfico actuales en dichas regiones (2). Estas diferencias deben ser consideradas para optimizar el diseño de paneles de marcadores para AM en poblaciones latinoamericanas.

En México históricamente los patrones genéticos resultado del mestizaje han sido profundamente influidos por las diferencias en las densidades de población parental y crecimiento demográfico debido a causas sociales y económicas (19-21). Si bien la heterogeneidad entre y al interior de los mestizos mexicanos de diferentes regiones del país se ha documentado mediante el uso de diversos marcadores genéticos (22-29), en la actualidad no existen en el dominio público análisis extensos de genoma completo para distintas poblaciones mestizas e indígenas en México.

Para analizar y caracterizar los patrones de diversidad genómica y desequilibrio de ligamiento en mexicanos esta trabajo incluye in-

Contribuciones de los autores: I.S. Z., A.H.M., J.E.G., C.L., y G.J.S. diseño de la investigación; I.S.Z., A.H.M., J.E.G., L.U.F., A.C., E.B.O., L.d.B.P., D.V.F., C.L., E.B., S.M., y G.J.S. desarrollo de la investigación; J.E.G. y C.D contribuyeron con nuevos reactivos/herramientas analíticas; y I.S.Z., A.H.M., J.E.G., J.C.F.L., L.U.F., R.G.,E.H.L., C.D., y G.J.S. analizaron los datos; y I.S.Z., A.H.M., J.E.G., y G.J.S escribieron el manuscrito.

Los autores declaran que no tienen conflicto de intereses.

Disponible gratuitamente en línea a través de la opción de PNAS acceso libre.

[1]I.S.-Z.,A.H.-M.,yJ.E.-G tuvieron la misma contribución a este trabajo.

[2]La correspondencia debe ser dirigida a: Gerardo Jiménez-Sánchez, MD, Ph.D. Instituto Nacional de Medicina Genómica, Periférico Sur No. 4124, Torre Zafiro II, 6to. Piso, Col. Jardines del Pedregal, México D.F. 01900, México. E-mail: gjimenez@inmegen.gob.mx

Este artículo contiene información suplementaria en línea: www.pnas.org/cgi/content/full/0903045106/DCSupplemental.

GENETICS

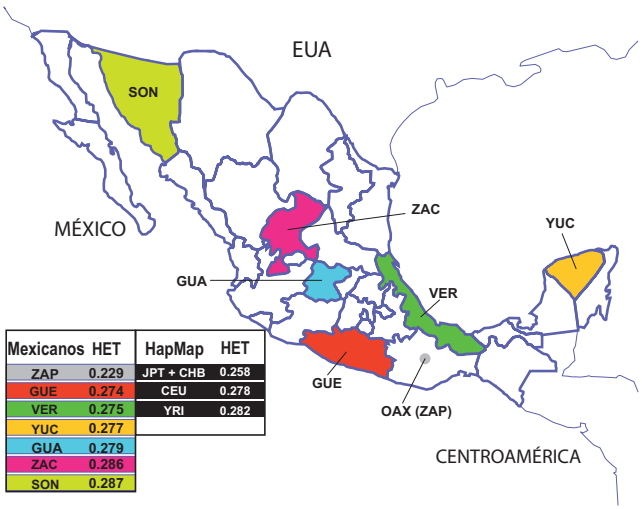

Fig. 1. Diversidad genética medida por heterocigocidad (HET) en mexicanos y poblaciones del HapMap. Los estados participantes se seleccionaron de tal forma que incluyeran las regiones Norte, Centro, Centro-Golfo, Centro-Pacífico y Sur de México. Se presentan valores promedio de heterocigocidad para amerindios zapoteca (ZAP); seis subpoblaciones mestizas mexicanas: Guanajuato (GUA), Guerrero (GUE), Sonora (SON), Veracruz (VER), Yucatán (YUC), y Zacatecas (ZAC); y poblaciones HapMap (YRI: africana, CEU: europea, y JPT+CHB: japonesa-china).

formación genotípica relativa 99,953 SNPs en 330 individuos provenientes de mestizos provenientes de seis regiones distintas y una población amerindia. Este estudio representa un primer esfuerzo en la construcción de una base de datos genéticos a nivel genómico que hemos denominado Proyecto de Diversidad Genómica de Mexicanos (MGDP por sus siglas en inglés). Este recurso será útil para desarrollar estrategias para el análisis genético de poblaciones mestizas, tales como la selección de marcadores para alcanzar una cobertura óptima de la variación genética común en estudios de asociación, tanto de genoma completo como de gen candidato. Así como para la adecuada aplicación de métodos de captura de diversidad genética o "tagging" y de imputación de genotipos (30, 31), y estrategias de mapeo genético por mestizaje (AM) (10) en mexicanos y otros poblaciones latinoamericanas. Nuestro estudio es uno de los primeros esfuerzos de genotipificación de genoma completo realizado en América Latina que pondrá a disposición del público estos datos. Este proyecto contribuirá al desarrollo de la medicina genómica tanto en México como en el resto de Latinoamérica.

Resultados

Analizamos datos de 300 sujetos autodefinidos mestizos no relacionados de 6 estados localizados en regiones geográficamente distantes de México. Estas regiones incluyeron Sonora (SON) y Zacatecas (ZAC) en el norte, Guanajuato (GUA) en el Centro, Guerrero (GUE) en el Centro-pacífico, Veracruz (VER) en la región Centro-Golfo, y Yucatán (YUC) en el Sureste (Fig. 1). Considerando que se ha demostrado que los amerindios Zapotecas son una población ancestral adecuada para estimar el origen ancestral amerindio en mestizos mexicanos (16), incluimos 30 zapotecas (ZAP) del estado sur occidental de Oaxaca en algunos de nuestros análisis. Para fines de comparación, también incluimos conjuntos de datos similares obtenidos de las poblaciones internacionales del HapMap: europeos del norte (CEU), africanos (YRI) y asiáticos orientales (AO), incluidos chinos (CHB) y japoneses (JPT). Se generó una base de datos con frecuencias de SNPs en mexicanos y poblaciones HapMap que puede ser consultada en http://diversity.inmegen.gob.mx.

Análisis de Diversidad Genética en mexicanos. Medimos la heterocigocidad (HET), realizamos análisis de componentes principales (PCA) (32), y calculamos estadísticos F_{ST} con conjuntos de datos obtenidos para las poblaciones mexicanas y HapMap. Las subpoblaciones mestizas mexicanas tuvieron valores de HET en un rango entre 0.274 en GUE y 0.287 en SON (Fig. 1). De los grupos HapMap los YRI tuvieron la más alta diversidad genética (HET= 0.282), en tanto que JPT+CHB, la más baja (HET=0.258), muy similar a valores reportados previamente (33). En mexicanos, las subpoblaciones del norte (SON y ZAC) mostraron los valores de HET más elevados sugiriendo mayor diversidad genética, y las muestras de amerindios ZAP tuvieron el valor más bajo (HET=0.229), tal como se espera para una población aislada (Fig. 1). Para el PCA, empleamos diferentes combinaciones de conjuntos de datos y condiciones. En todos los escenarios se exhiben los vectores Eigen más informativos para cada conjunto de datos (Fig. 2A-D). Cuando se incluyen, las poblaciones del HapMap y los ZAP forman grupos claramente definidos, en tanto que las subpoblaciones mestizas mexicanas se distribuyen ampliamente entre las muestras CEU y ZAP (Fig. 2A y 2B). El agrupamiento de la población ZAP en el gráfico de PCA sugiere que los individuos amerindios incluidos en nuestro estudio no muestran evidencia de mestizaje reciente. Como era de esperarse, cuando se analizaron todos los grupos (Fig. 2A), el primer eje mostró que la mayor distancia genética se presenta entre la población africana (YRI) y el resto de los grupos. En el segundo eje, el grupo ZAP se localiza entre CEU y los asiáticos orientales (CHB y JPT) y, tanto en el primero como el segundo eje, los grupos mestizos mexicanos muestran una distribución amplia entre CEU y ZAP (Fig. 2A y 2B). Para visualizar mejor la distribución de mestizos mexicanos, generamos dos conjuntos de datos adicionales, uno en el que se excluyeron las muestras YRI (Fig. 2B) y un segundo incluyendo sólo las dos poblaciones que resultaron genéticamente más cercanas a los mestizos mexicanos: CEU y ZAP (Fig. 2C). Estos análisis evidenciaron la presencia de diversidad genética entre y dentro de las poblaciones de mestizos mexicanos. Además, un PCA que incluyó sólo a CEU, ZAP y las dos subpoblaciones mestizas con las mayores diferencias en valores de HET (SON y GUE), reveló que la mayoría de las muestras de SON se localizan más cerca a CEU, y la mayoría de las de GUE al grupo ZAP (Fig. 2C y 2D). En ambas Fig.s, algunos individuos se desplazaron a lo largo del Eigen vector 2, lo que refleja contribuciones ancestrales adicionales a los mestizos. Para evaluar si este efecto está relacionado con la contribución ancestral africana, analizamos un conjunto adicional de datos que incluyó a los YRI (Fig. suplementaria 1A y 1B). La distribución de mestizos en el Eige vector 3 (Fig. suplementaria 1B) indica que la dispersión observada en las Fig.s 2D y 2C en el Eigen vector 2 reflejan una contribución ancestral africana a los mestizos. Resulta interesante que en este PCA, los mestizos no se organizan en una línea recta entre CEU y ZAP. Esto puede deberse a que ambos grupos no representan en su totalidad a la variabilidad genética de orígenes europeo y amerindio en estos mestizos (2). Para medir las diferencias genéticas entre las subpoblaciones mexicanas, así como entre éstas y los grupos HapMap, realizamos un análisis estadístico F_{ST} pareado (Tabla 1). De todos los grupos mexicanos, la población amerindia ZAP mostró los valores F_{ST} más altos cuando se comparó con todas las poblaciones HapMap. Como era de esperarse, el valor superior se observó cuando se comparó con la YRI (23.9), seguido por la CEU (15.4), JPT (11.9) y CHB (12.0). Los valores F_{ST} entre ZAP y cada una de las subpoblaciones mestizas (Tabla 1) fueron consistentes con su distribución en el gráfico del PCA (Fig. 2C), siendo GUE y VER los más cercanos al grupo ZAP (valores F_{ST} 3.2 y 3.8, respectivamente) y SON en el otro extremo de la distribución (F_{ST} de 8.2). Las comparaciones en pares entre los grupos mexicanos reveló que SON al compararse

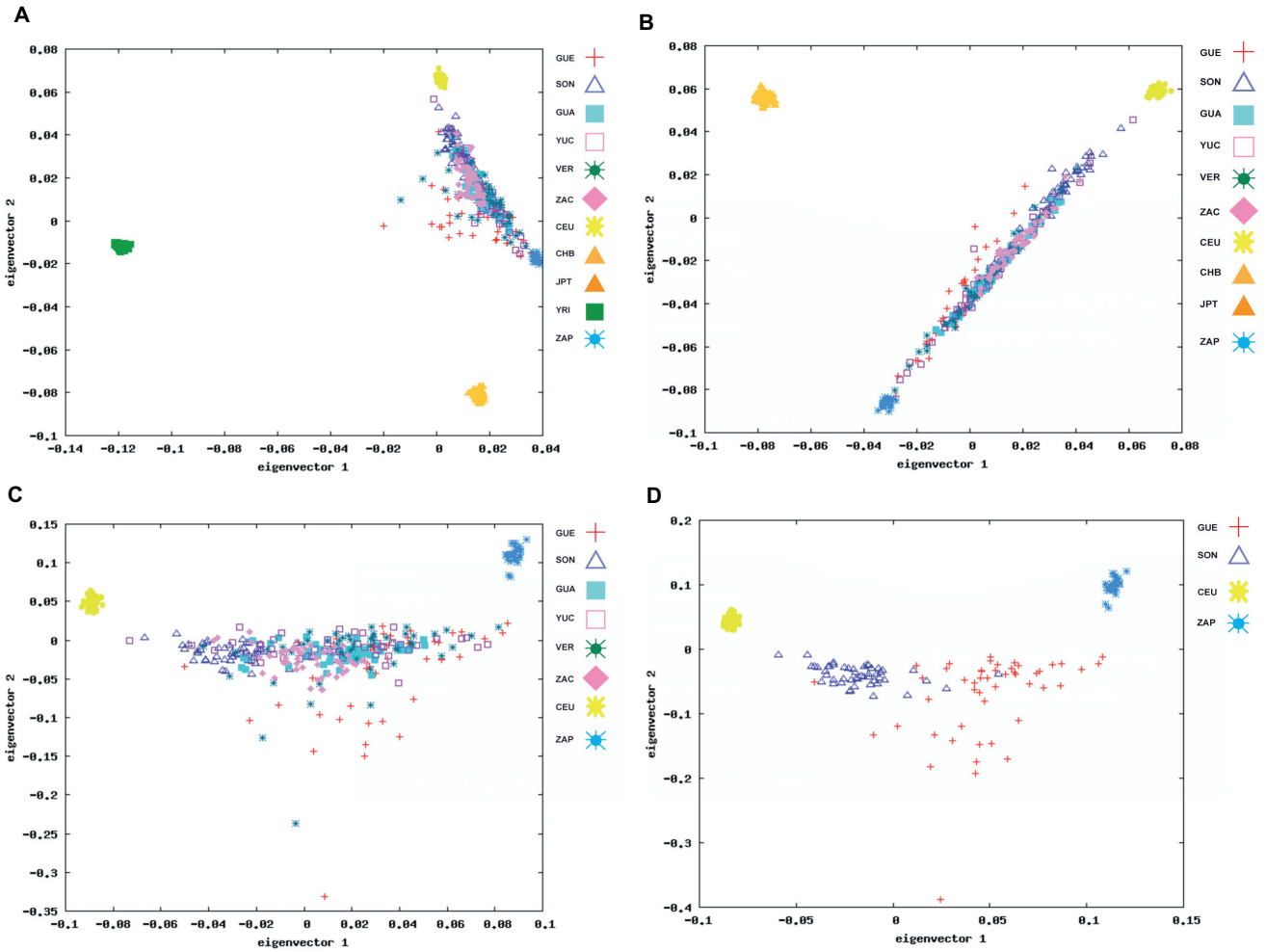

Fig. 2. Análisis de componentes principales. Los dos Eigen vectores más informativos se graficaron en todos los casos. Se presentan cuatro conjuntos de datos diferentes: A) todas las subpoblaciones mexicanas, tanto mestizas (GUA, GUE, SON, VER, YUC, ZAC) y amerindios (ZAP), así como poblaciones HapMap (YRI, CEU y JPT+CHB); B) todos los mestizos mexicanos, ZAP, CEU, y JPT+CHB; C) todos los mestizos mexicanos, ZAP y CEU; y D) las dos subpoblaciones mestizas mexicanas que muestran las mayores diferencias en el Eigen vector 1 (SON y GUE), ZAP y CEU.

con todos las grupos mestizos presenta valores de F_{ST} más elevados que los observados entre los CHB y JPT. Mas aún, el valor F_{ST} entre SON y ZAP (8.2) fue mayor que cualquier otro resultado de comparación entre subpoblaciones mestizas y grupos no-africanos del HapMap (Tabla 1). Estos resultados apoyan la presencia de una considerable heterogeneidad genética entre las subpoblaciones mestizas mexicanas de regiones geográficamente distantes de México, y sugiere que las diferencias observadas se asocian con una distribución diferencia de los componentes de ancestrales, en particular amerindio (AMI) y europeo (EUR).

Tabla 1. Valores FST entre mexicanos, amerindios zapoteca y poblaciones HapMap.

	GUE	SON	VER	YUC	ZAC	ZAP	CEU	YRI	CHB	JPT
GUA	0.2	1.1	0.1	0.3	0.1	4.3	5.2	15.4	6.9	6.9
GUE		1.9	0.1	0.4	0.5	3.2	6.9	15.7	7.0	7.0
SON			1.3	1.2	0.6	8.2	2.0	13.9	7.3	7.4
VER				0.2	0.2	3.8	5.8	15.7	6.9	7.0
YUC					0.3	4.5	5.2	15.6	7.0	7.0
ZAC						5.3	4.0	14.5	6.8	6.9
ZAP							15.4	23.9	12.0	11.9
CEU								15.7	11.0	11.2
YRI									18.4	18.5
CHB										0.7

Se realizaron cálculos utilizando el conjunto completo de SNPs (99,953) con el programa EIGENSOFT.

Para evaluar los componentes genéticos ancestrales en los mexicanos, se determinaron proporciones ancestrales individuales y poblacionales promedio, utilizando STRUCTURE (34, 35). Para dichos análisis, utilizamos un conjunto de 1,814 marcadores informativos de origen ancestral (AIMs, por sus siglas en inglés) seleccionados utilizando diferentes criterios para garantizar cobertura a lo largo de todo el genoma y minimizar el LD entre SNPs (Ver Materiales y Métodos). Utilizamos a las muestras del HapMap y la población ZAP como poblaciones ancestrales europea (EUR), africana (AFR), asiática oriental (EA) y amerindia (AMI), respectivamente. Nuestros resultados fueron consistentes con cuatro contribuciones poblacionales (K=4), que explican la subestructura principal presente en el conjunto de mestizos mexicanos analizados (Fig. 3A y 3B). En este modelo, los promedios ancestrales en mestizos mexicanos fueron: 0.552±0.154 para AMI, 0.418±0.155 para EUR, 0.018±0.035 para AFR, y 0.012±0.018 para EA (Tabla 1 suplementaria). Observamos diferencias entre y al interior de las subpoblaciones mestizas mexicanas, en particular en las contribuciones ancestrales EUR y AMI (Fig. 3A y 3B). Las estimaciones más altas y más bajas de contribución ancestral promedio EUR fueron 0.616±0.085 para SON y 0.285±0.120 para GUE. La mayor parte de las subpoblaciones mestizas exhibieron diferencias estadísticamente significativas en la contribución ancestral prome-

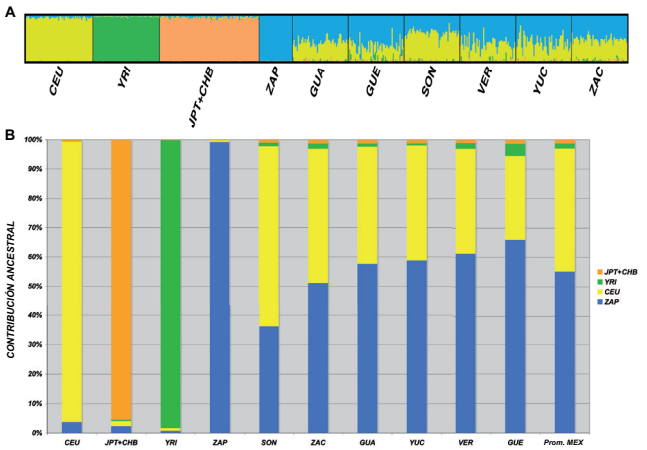

Fig. 3. Análisis de estructura poblacional utilizando 1,814 AIMs. A) Proporciones ancestrales individual. Cada línea vertical representa un individuo en una población dada. Las contribuciones de poblaciones ancestrales tienen un código de color: amarillo=europeo (EUR), verde=africano (AFR), naranja=asiático oriental (EA), y azul=amerindio (AMI) para seis subpoblaciones mestizas mexicanas: Guanajuato (GUA), Guerrero (GUE), Sonora (SON), Veracruz (VER), Yucatán (YUC), y Zacatecas (ZAC). B) Contribuciones ancestrales promedio en subpoblaciones mestizas mexicanas. Se observaron diferencias significativas en proporciones ancestrales principalmente para contribuciones europeas y amerindias (Tabla 2 suplementaria).

dio EUR, y tanto SON como GUE mostraron diferencias cuando se compararon con cualquiera de las otras subpoblaciones mestizas (Tabla 2 suplementaria). Los únicos grupos con contribución ancestral promedio EUR similar, fueron aquellos procedentes de la región central y centro-costera (VER, YUC, y GUA). En contraste, la mayoría de las subpoblaciones mestizas analizadas tuvieron valores similares de contribución ancestral promedio AMI —GUE mostrando las mayor contribución (0.660±0.138), y SON la menor (0.362±0.089) (Fig. 3B)— y sólo las subpoblaciones en los estados norteños (SON y ZAC) mostraron diferencias estadís-

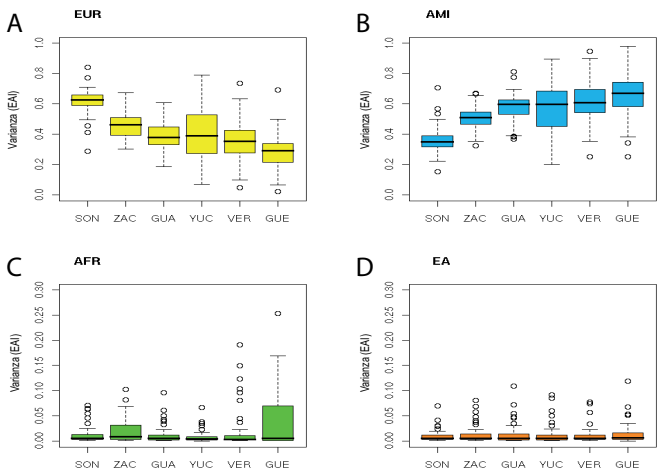

Fig. 4. Diagramas de cajas y bigotes de estimaciones de componentes ancestrales. Distribución por cuartiles de proporciones ancestrales para seis subpoblaciones mestizas mexicanas: Guanajuato (GUA), Guerrero (GUE), Sonora (SON), Veracruz (VER), Yucatán (YUC), y Zacatecas (ZAC). Los paneles corresponden a cada una de las poblaciones parental: A) europeos, B) amerindios, C) africanos, y D) asiático-orientales. La gráfica representa los valores mínimos y máximos (bigotes), el primero y tercer cuartil (caja), y el valor medio (línea media). También se muestran los valores atípicos. El eje de las Y representa la varianza del estimado ancestral individual (EAI) da acuerdo con el cálculo de STRUCTURE.

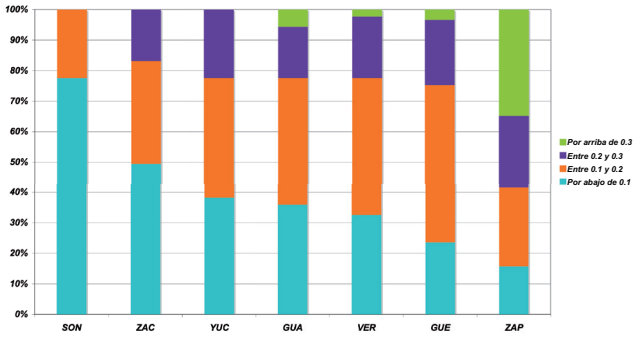

Fig. 5. Distribución de frecuencias de SNPs privados para mexicanos comparado con poblaciones HapMap. Se definieron los SNPs privados como aquellos que presentan MAF>0.05 en al menos una subpoblación mexicana, pero que está ausente en todas las poblaciones HapMap. Cada barra representa la distribución de frecuencias de todos los SNPs privados (n=86) para cada subpoblación mexicana.

ticamente significativas cuando se compararon con el resto de los grupos mestizos (Tabla 2 suplementaria). Las otras dos contribuciones ancestrales incluidas en nuestro análisis, AFR y EA, fueron más pequeñas y casi homogéneas entre todas las subpoblaciones mestizas. Se observaron diferencias significativas en el componente ancestral AFR para SON y ZAC comparadas con VER y YUC (Tabla 2 suplementaria). En nuestro análisis no se observaron diferencias significativas en las estimaciones de EA (Tabla 2 suplementaria). Para evaluar la contribución de las diferencias en componentes ancestrales a la diversidad genética regional global entre subpoblaciones mestizas, calculamos los coeficientes de correlación de Pearson entre valores pareados de F_{ST} y las diferencias en las proporciones ancestrales en AMI, EUR y AFR. Este análisis nos permitió identificar una alta correlación entre la diversidad genética global (F_{ST}) y las diferencias en componentes ancestrales EUR (r=0.937) y AMI (r=0.944). Para estimar la magnitud de este efecto, calculamos la distancia genética entre las subpoblaciones mexicanas específicamente atribuible a las diferencias en las dos principales proporciones de origen ancestral continental (Tabla 3 suplementaria). Este análisis reveló que para la mayoría de las comparaciones entre pares de subpoblaciones mestizas (10 de 15), más del 50% de la distancia genética entre ellas es atribuible a las diferencias en el origen ancestral continental. Resulta interesante que la subpoblación de YUC estuvo presente en la mayoría de las comparaciones que exhiben los valores más bajos, de contribución de las diferencias en componentes ancestrales continentales a la distancia genética global. El grupo de YUC son los únicos mestizos en este estudio con un componente ancestral AMI principal distinto al resto, la contribución Maya. Estos resultados sugieren que la estructura poblacional en los mestizos mexicanos está relacionada con diferencias en las contribuciones de ancestral continental, aun cuando haya sido afectada por otras fuentes de diversidad genética, tales como contribuciones amerindias características.

Para evaluar la magnitud de las diferencias intra-regionales en proporciones ancestrales continentales entre mestizos mexicanos, comparamos las distribuciones para cada contribución ancestral usando diagramas de cajas y bigotes (Fig. 4A, B, C y D) y se midieron los coeficientes de variación (CVs) (Tabla 4 suplementaria), como una medida normalizada de la dispersión de cada componente. En los CVs observamos una amplia distribución, en un rango de 0.139-0.421 para EUR, de 0.151-0.273 para AMI, de1.236-2.096 para AFR, y de 1.264-1.625 para EA. Para las contribuciones ancestrales EUR y AMI se observó una varianza baja en todas las subpoblaciones, y en GUE y YUC se observó la mayor dispersión para EUR y AMI, 0.421 y 0.273 respectivamente (Fig.

Tabla 2. Porcentaje de haplotipos comunes compartidos entre las poblaciones Mexicana y HapMap.

Población	CEU	JPT + CHB	YRI	CEU +JPT + CHB	CEU + JPT + CHB + YRI
GUA	81	75	64	93	96
GUE	79	76	65	93	96
SON	82	71	63	94	97
VER	80	75	64	93	96
YUC	81	74	64	93	96
ZAC	81	73	64	93	97
MEX Prom	81	74	64	93	96

Se evaluó la compartición de haplotipos mediante la comparación de frecuencias de 5 extensiones de haplotipos SNP consecutivos de ~100kb. La tabla muestra el porcentaje de haplotipos comunes compartidos (frecuencia >5%) entre las poblaciones mexicana y HapMap.

4A y B). En los grupos de VER y GUE se identificó la presencia de individuos con valores atípicos de componente AFR, con contribuciones ancestrales superiores a 15% y la mayor variabilidad intra-regional en este componente (CVs=2.096 y 1.501) (Fig. 4C). Aun cuando las contribuciones EA estimadas fueron las más pequeñas, se observó una dispersión alta (CVs=1.264-1.625) para todas las subpoblaciones (Fig. 4D). Estos resultados sugieren que la estructura poblacional en mestizos mexicanos está principalmente relacionada a diferencias en proporciones ancestrales EUR y AMI, pero que otras fuentes de diversidad genética, tales como las contribuciones AFR y particulares fuentes AMI contribuyen también a las diferencias genéticas observadas.

Alelos privados en poblaciones mexicanas. Para identificar variaciones genéticas específicas de mexicanos con respecto a las poblaciones HapMap se buscaron alelos privados. Se identificaron 89 alelos privados comunes con frecuencias de alelo menor (MAF, por sus siglas en inglés) >0.05 ausentes en las poblaciones HapMap, pero presentes en al menos una sub-población mestiza mexicana, y 86 alelos privados de amerindios mexicanos (ZAP). Todos los alelos privados para ZAP también lo fueron para los mestizos, lo que señala el origen amerindio de los mismos. El número de alelos privados fue similar en los seis estados, pero se observaron diferencias en la proporción de variantes con frecuencias superiores al 20% (MAF>0.20). No observamos alelos con MAF>0.20 en SON, ni con MAF>0.30 en ZAC o YUC (Fig. 5). Estos resultados soportan la observación de que, de las subpoblaciones analizadas, las del norte de México (SON y ZAC) tienen la mayor contribución ancestral EUR, y las de la región centro-costera (GUE y VER) tienen el mayor componente AMI. Para analizar este resultado en el contexto de contribuciones genéticas continentales buscamos alelos privados para cada grupo HapMap comparado con el resto e identificamos 5660 alelos privados para YRI, 1533 para CEU, y 669 para CHB+JPT. La observación del mayor número de alelos privados en africanos y el menor en amerindios es consistente con lo esperado de acuerdo a los modelos de evolución humana que señalan el origen en África y la subsiguiente llegada de los seres humanos a América después de una serie de efectos fundadores (36).

Tabla 3. Porcentaje de haplotipos comunes compartidos entre las subpoblaciones mexicanas.

Población	GUA	GUE	SON	VER	YUC	ZAC
GUA		86	87	87	87	88
GUE	88		87	88	87	88
SON	83	82		82	83	84
VER	87	86	87		87	88
YUC	87	85	87	86		87
ZAC	85	83	87	85	85	
MEX Prom	86	84	87	86	86	87

Se evaluó el porcentaje de haplotipos compartidos mediante la comparación de frecuencias haplotipos extendidos de 5 SNPs consecutivos de ~100kb. La tabla muestra el porcentaje de haplotipos comunes compartidos (frecuencia >5%) entre subpoblaciones mexicanas.

Para identificar las regiones genómicas con diferencias intra-poblaciones en México, buscamos alelos presentes en una subpoblación mestiza mexicana particular, pero ausente en las otras cinco. En este análisis encontramos sólo dos SNPs con frecuencias >0.05, uno en SON (rs5973601, MAF=0.053) y uno en ZAC (rs3733654, MAF=0.051). El análisis de grado de informativo para asignación o "informativeness for assignment" (37) se utilizó como un acercamiento para intentar identificar un subconjunto de SNPs presentes en la mayoría de subpoblaciones mexicanas, pero con variación geográfica en su frecuencia alélica. Este análisis dio lugar a la identificación de 14 SNPs con el contenido de información más alto (In >0.04) (Fig. suplementaria 2). Todos los SNPs en este conjunto resultaron también ser AIMs con $\delta \geq 0.27$ para al menos uno de los orígenes ancestrales incluidos en análisis previos (Tabla suplementaria 5). Este resultado proporciona apoya el que las diferencias genéticas observadas entre las diferentes regiones en México están asociadas a las contribuciones ancestrales continentales, y destaca regiones genómicas con diferencias intra-poblacionales importantes en la frecuencia alélica de algunos SNPs. Estas regiones pudieran ser fuente de señales falso-positivas en estudios de asociación genética en mexicanos.

Patrones de desequilibrio de ligamiento (LD) en poblaciones mestizas mexicanas y poblaciones HapMap. La distribución promedio de frecuencia de SNPs comunes (MAF >15%) fue similar en las muestras mexicanas que en las poblaciones del HapMap (Fig. Suplementaria 3A), indicando la ausencia de sesgo de información o determinación. SON y ZAC tuvieron una proporción menor de marcadores de frecuencia baja (MAF<5%) que las poblaciones HapMap para lo cual señala un grado menor de homocigocidad en estos grupos. Este resultado es consistente con el hecho de que SON y ZAC son los grupos mestizos con las HET más altas (Fig. 1). Para evaluar el tamaño potencial de los bloques de haplotipos en mestizos mexicanos, comparamos los gráficos de decaimiento de LD de alelos comunes (MAF >15%) altamente correlacionados ($r2 > 0.8$ y ($|D'| > 0.8$) entre las poblaciones mexicanas y las del HapMap. El decaimiento de LD en mexicanos fue similar al de las muestras HapMap no africanas (Fig. Suplementarias 3B y C). Para evaluar con más detalle la variabilidad de la estructura genómica en poblaciones mexicanas y HapMap, realizamos análisis de diversidad de haplotipos extendidos (LRHD por sus siglas en inglés). Cuando las poblaciones mexicanas se compararon con las del HapMap, la mayoría presentó menor diversidad, y sólo SON presentó un patrón de LRHD similar a los de asiáticos orientales (CHB y JPT). De todas las subpoblaciones mexicanas, GUE mostró la menor diversidad haplotípica y SON la mayor (Fig. suplementaria 4A). En promedio, 68 haplotipos por Mb representan 95% de los cromosomas en muestras mexicanas, en tanto que la misma cobertura requirió 93, 83, 69, y 70 haplotipos en las muestras YRI, CEU, CHB, y JPT, respectivamente (Fig. Suplementaria 4B). Este resultado apunta hacia una menor diversidad de haplotipos en poblaciones mexicanas que en poblaciones HapMap.

Análisis de haplotipos compartidos entre poblaciones mestizas mexicanas y las del HapMap. Para determinar el uso potencial de los datos HapMap para estudios de asociación genética en regiones candidato y genoma completo en mexicanos, evaluamos el número de haplotipos con una frecuencia compartida >5% entre las poblaciones mexicanas y HapMap. Este análisis mostró que los mexicanos comparten 64% de estos haplotipos con YRI, 74% con JPT+CHB,

y 80% con CEU. La proporción de haplotipos compartidos aumenta a 96% cuando se comparan con la combinación de las cuatro poblaciones HapMap (Tabla 2). Estos resultados muestran que la cobertura efectiva de variaciones genéticas comunes en mexicanos es factible mediante el uso de la información HapMap. No obstante, lo anterior podría estar asociado a un elevado costo de genotipificación debido a la necesidad de incluir el conjunto de datos combinados para todas las poblaciones HapMap. Para evaluar el beneficio potencial de utilizar un mapa de haplotipos diseñado en base a información obtenida de población mexicana, en lugar de sólo la información HapMap, evaluamos los haplotipos compartidos entre subpoblaciones mexicanas. Este análisis incluyó comparaciones para las cuales se utilizó una subpoblación mexicana como referencia para las otras, como se hizo previamente para cada grupo del HapMap. Llevamos a cabo el mismo análisis utilizando todos los posibles pares de subpoblaciones mexicanas como el grupo de referencia. Este análisis reveló que todas las subpoblaciones mexicanas comparten, en promedio, 86% (84-87%) de los haplotipos comunes cuando se utiliza como referencia una de ellas (Tabla 3), y que la proporción de haplotipos compartidos aumenta a un promedio de 96% (95-97%) cuando cada subpoblación se compara con cualquiera de los pares de subpoblaciones (Tabla suplementaria 6). Estos resultados apoyan la idea que la generación del mapa de haplotipos de la población mestiza mexicana puede contribuir a reducir el número de SNPs marca o "tag" requerido para capturar y caracterizar variaciones genéticas comunes en la población.

Discusión

Este trabajo es una evaluación inicial del beneficio potencial de generar un mapa de haplotipos de mexicanos para optimizar el diseño y análisis de estudios de asociación genética en esta población. Durante el periodo pre-hispánico, eran más numerosos los grupos étnicos que vivían en el centro y sur de México y contaban con una cohesión política, religiosa y social más sólida que los grupos étnicos de la región del norte. Entre 1545 y 1548 se trajeron esclavos negros a la región después del notable descenso de la población amerindia causada por las epidemias (19). Desde entonces, los procesos de mestizaje en regiones geográficamente distantes se han visto afectadas por diferentes condiciones demográficas e históricas que han modelado la estructura genómica de los mexicanos. Estos factores han generado heterogeneidad genética entre y al interior de las diferentes regiones de todo México (2, 26, 29, 38). A pesar de que la muestra de participantes en nuestro estudio se obtuvo de regiones que corresponden a subdivisiones políticas actuales, ésta representa ejemplos de las diferentes dinámicas demográficas, patrones de asentamientos humanos y densidades de poblaciones amerindias. Dado que estudios previos han descrito sesgos en estimados de mezclas para mexicanos debido a la estratificación socioeconómica (28), se reclutaron participantes mestizos de universidades estatales, en las que la mayoría de los asistentes proviene de áreas tanto urbanas como rurales, y pertenece a una amplia gama de estratos socio-económicos.

Nuestros resultados muestran que las diferencias genéticas entre mestizos de diferentes regiones de México, se deben principalmente a diferencias en contribuciones ancestrales EUR y AMI. En la mayoría de los análisis, las muestras de las regiones centrales se comportaron cercanas a los ZAP, mientras que las muestras de regiones del norte se comportaron cercanas a los CEU, correlacionando con la densidad poblacional amerindia actual y pre-hispánica en esas regiones (19). Aunque nuestro análisis mostró que la contribución ancestral AFR fue baja (<10%) y en su mayoría homogénea entre las subpoblaciones, sí observamos la presencia de individuos con una contribución ancestral AFR particularmente alta en GUE y VER. Esto se apega a los registros históricos que señalan a estos estados como el principal punto de entrada de africanos durante el periodo colonial, y como residencia de afro-mexicanos desde entonces (39). Resulta interesante, que las muestras de la región sureste (YUC) muestran la menor contribución de las diferencias de componentes ancestrales continentales a la distancia genética global. Los mestizos de Yucatán constituyen el único grupo en nuestra muestra cuya contribución amerindia es principalmente maya. Los Mayas representan un grupo étnico geográficamente distante de otros grupos amerindios mexicanos, y con marcadas diferencias culturales, sociales e históricas respecto a estos (20). Este resultado sugiere que parte de la diversidad genética observada en estos mestizos se asocia a contribuciones amerindias diferenciales.

Los alelos privados de mestizos mexicanos tienen un origen amerindio y constituyen una representación conservadora de la variación genética que está ausente en otros grupos continentales, sobretodo si consideramos que la mayoría de los SNPs analizados fueron identificados en poblaciones sin una contribución genética ancestral de origen amerindio (40). La detección exitosa de SNPs privados AMI está relacionada al uso de un ensayo de genotipificación con información de un grupo multiétnico que incluyó hispanos/latinos del DNA Polymorphism Discovery Resource (http://www.genome.gov/10001552) (40). Estos SNPs representan variantes que no has sido cubiertas en el HapMap y que podrían no ser capturadas si SNPs marca o "tag" se seleccionan utilizando sólo información de estos grupos poblacionales. Para describir mejor los SNPs y haplotipos privados de mexicanos y sus SNPs marca o "tag" asociados, es necesario llevar a cabo amplios proyectos de re-secuenciación que consideren tanto mestizos como amerindios mexicanos.

Considerando la similitud de los patrones de decaimiento de LD de las poblaciones mexicana con los de todas las poblaciones no africanas del HapMap, se espera que el tamaño de bloques de haplotipos promedio en mexicanos sea similar al de las poblaciones europeas y asiáticas. Los niveles bajos de LRHD observado en mexicanos comparado con las poblaciones HapMap correlaciona con la contribución AMI, y es consistente con el hecho de que las poblaciones amerindias presentan una diversidad haplotípica reducida y patrones de LD de largo alcance (35), y por lo tanto puede estar relacionado a la disminución progresiva en la diversidad haplotípica en poblaciones humanas que emigran desde África (41). El análisis de haplotipos compartidos se utilizó como abordaje para estimar en forma indirecta la transferabilidad de SNPs marca o "tag" de las poblaciones HapMap a las mexicanas, y entre las subpoblaciones mexicanas. Este análisis se realizó utilizando diferentes combinaciones de poblaciones mexicanas y HapMap para evaluar los beneficios potenciales de un mapa mexicano de haplotipos. La variación genética común en los mexicanos se cubre de manera efectiva (96%) únicamente cuando se utilizan los datos combinados de todas las poblaciones HapMap, de acuerdo con hallazgos previos para poblaciones latinas (12). Esto sugiere que la selección de los SNPs marca o "tag" utilizando sólo información HapMap, daría lugar a costos más elevados en estudios de asociación en población mexicana derivados de una sobre-genotipificación. Una indicación de que el mapa de haplotipos para mexicanos podría ser útil para la selección de SNPs marca o "tag" es que el uso de cualquier combinación de dos subpoblaciones mexicanas como referencia logra una mejor cobertura que el uso de la combinación de todas las poblaciones HapMap. Estos resultados apoyan el hecho de que un mapa de haplotipos que describa en forma integral la variabilidad genética común y los patrones LD en mexicanos es factible y útil.

La disponibilidad pública de los datos del MGDP será de gran importancia para hacer un más eficiente diseño de estudios de asociación y proyectos de re-secuenciación en poblaciones latinas. Nuestro estudio sugiere que tanto los estudios de asociación de

genoma completo como aquellos de región y/o gen candidato basados en el uso de SNPs marca o "tag" de los cuatro grupos HapMap captura adecuadamente 96% de la diversidad genética común en mexicanos. Sin embargo, parece posible generar conjuntos óptimos de SNPs marca para mexicanos, para mejorar la eficiencia en la captura de señales de asociación en estudios de región y/o gen, y contribuir así a reducir costos sin comprometer la cobertura. Esto es crítico en México y otros países de América Latina en los cuales el financiamiento para investigación a menudo es limitado. Además, disponer de un mapa de haplotipos de la población mexicana podría ser útil para mejorar la captura o "tagging" de haplotipos, mejorando así las estrategias para el descubrimiento de SNPs en mexicanos. Lo anterior, un asunto de máxima importancia en la investigación y búsqueda de variantes raras asociadas con enfermedades comunes complejas.

Los métodos de imputación de genotipos faltantes o inciertos se han aplicado con éxito para mejorar el poder y combinar datos estudios de asociación de genoma completo (30, 31). Sin embargo, este abordaje asume que existen patrones similares de LD entre las muestras analizadas y el panel de referencia (30). La captura o "tagging", así como la imputación de alelos en mexicanos u otras poblaciones latinas con una historia demográfica similar utilizando información HapMap, podría no ser del todo precisa por la presencia de un componente genético no capturado con los datos HapMap (13). La disponibilidad pública del MGDP proporcionará un conjunto de datos de gran valor para el desarrollo de métodos que permitan examinar la precisión del paradigma de "imputación" en una población de reciente mestizaje, así como para optimizarlos utilizando información acerca de la composición ancestral de los individuos en la población. Los datos del MGDP también serán útiles para optimizar los conjuntos de AIMs descritos para los estudios de AM en poblaciones latinas (14-16, 29), favoreciendo diseños más robustos en rasgos y enfermedades que muestran diferencias de prevalencia por etnicidad en mexicanos, como son los niveles de colesterol HDL (42), las enfermedades de la vesícula biliar (43), y la diabetes tipo 2 (44).

Con objeto de cubrir en forma integral la variabilidad genética común y describir mejor la estructura genómica de los mexicanos, estamos aumentando la densidad de SNPs a ~1.5 millones por genoma, para lo cual se está utilizando una combinación de plataformas de microarreglos. Aquí presentamos uno de los primeros conjunto de datos de genoma completo para poblaciones mexicanas mestizas y amerindias. Este esfuerzo contribuirá al diseño de mejores estrategias encaminadas a caracterizar y comprender mejor los factores genéticos subyacentes a las enfermedades comunes complejas de los mexicanos. Además, esta información incrementará nuestro conocimiento sobre la variabilidad genómica de las poblaciones de América Latina. La infraestructura científica y tecnológica derivada de este proyecto favorecerá el desarrollo de la medicina genómica en México y América Latina (3, 6).

Materiales y Métodos

Se colectaron 300 muestras de sangre de individuos no emparentados autodefinidos como mestizos y de 30 amerindios zapotecas sin parentesco familiar en primer grado. Se obtuvieron muestras anónimas en siete estados de México: Guanajuato. Guerrero, Sonora, Veracruz, Yucatán, Zacatecas, y Oaxaca (ZAP). Este estudio fue aprobado por los Comités Científico, Ético y de Bioseguridad del Instituto Nacional de Medicina Genómica (INMEGEN). Se aplicó un proceso ad-hoc para la consulta y participación comunitaria que incluyó tanto a autoridades gubernamentales estatales y autoridades universitarias y sanitarias, como a miembros de la comunidad local. Se realizó la extracción de DNA genómico de sangre periférica (Qiagen). La genotipificación se realizó conforme al protocolo del microarreglo Affymetrix 100K y en total, 99,953 SNPs cumplieron con los criterios de control de calidad para todas las poblaciones analizadas. El ordenamiento o "phasing" de nuestros genotipos mexicanos se realizó con fastPhase v1.1.4 (45). Existe un portal de internet (http://diversity.inmegen.gob.mx) en el que se puede tener acceso a la base de datos tipo HapMap que incorpora frecuencias de SNPs de poblaciones mestizas mexicanas, Zapotecos y HapMap. Todos los datos de genotipos y señales crudas de intensidad están disponibles para su descarga en el sitio (ftp://ftp.inmegen.gob.mx).La heterocigocidad (HET) promedio se calculó con PLINK http://pngu.mgh.harvard.edu/purcell/plink/ (46). Los análisis de componentes principales (PCA por sus siglas en inglés) se realizaron con el programa EIGENSTRAT (32) y los de con EIGENSOFT (39). Se utilizó el programa STRUCTURE v.2.1 (34, 35), incluyendo 1,814 marcadores Informativos de origen ancestral (AIMs, por sus siglas en inglés) (Ver Materiales y Métodos Suplementarios). N. Rosenberg amablemente proporcionó el programa para calcular el grado de informativo para asignar o "informativeness for assignment" de los marcadores (37). Los alelos privados a la población Mexicana se definieron como SNPs presentes en cualquiera de las subpoblaciones mexicanas, pero ausentes en todas las poblaciones HapMap. Para ello, consideramos SNPs con un MAF>0.05. Para identificar alelos privados para cualquier sub-población mexicana específica, se llevó a cabo un análisis similar (pero sin datos HapMap) en el cual investigamos SNPs presentes en un grupo mexicano y ausentes en los otros seis. Los cálculos de LD, diversidad de haplotipos extendidos (LRHD, por sus siglas en inglés), y análisis de haplotipos compartidos se realizaron con Haploview y programas específicamente generados para ese fin, tal como se describió previamente (47, 48) . Todos los análisis fueron realizados en el INMEGEN en la Ciudad de México. (Ver Materiales y Métodos Suplementarios).

AGRADECIMIENTOS. Agradecemos al Gobierno Federal de México, en particular a la Secretaría de Salud, por su valioso apoyo en todas las etapas de este proyecto. La participación de los Gobiernos y Universidades de los estados de Guanajuato, Guerrero, Oaxaca, Sonora, Veracruz, Yucatán y Zacatecas contribuyó en forma significativa a este trabajo. Agradecemos a todos los voluntarios de este estudio, en particular a aquellos que contribuyeron con sus muestras sanguíneas. A Alejandro López, José Bedolla y Lucía Orozco por sus importantes contribuciones a través de la estrategia de comunicación. A la Dra. Blanca Z. González Sobrino por su contribución a la discusión acerca de la etnohistoria de México. Este trabajo recibió apoyo financiero del Gobierno Federal de México canalizado al Instituto Nacional de Medicina Genómica, y la infraestructura fue donada por la Fundación Mexicana para la Salud (FUNSALUD) y la Fundación Gonzalo Río Arronte.

1. Gonzalez Burchard E, et al. (2005) Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. *Am J Public Health* **95,** 2161-2168.
2. Wang S, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* **4,** e1000037.
3. Jimenez-Sanchez G (2003) Developing a platform for genomic medicine in Mexico. *Science* **300,** 295-296.
4. Hardy BJ, et al. (2008) The next steps for genomic medicine: challenges and opportunities for the developing world. *Nat Rev Genet* **9** Suppl 1, S23-27.
5. Seguin B, Hardy BJ, Singer PA, & Daar AS (2008) Genomics, public health and developing countries: the case of the Mexican National Institute of Genomic Medicine (INMEGEN). *Nat Rev Genet* **9** Suppl 1, S5-9.
6. Jimenez-Sanchez G, Silva-Zolezzi I, Hidalgo A, & March S (2008) Genomic medicine in Mexico: Initial steps and the road ahead. *Genome Res* **18,** 1191-1198.
7. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661-678.
8. McCarthy MI, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9,** 356-369.
9. Smith MW & O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat Rev Genet* **6,** 623-632.
10. Seldin MF (2007) Admixture mapping as a tool in gene discovery. *Curr Opin Genet Dev* **17,** 177-181.
11. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437,** 1299-1320.
12. de Bakker PI, et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* **38,** 1298-1303.
13. Huang L, et al. (2009) Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84,** 235-250.
14. Mao X, et al. (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. *Am J Hum Genet* **80,** 1171-1178.
15. Tian C, et al. (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am J Hum Genet* **80,** 1014-1023.
16. Price AL, et al. (2007) A genomewide admixture map for Latino populations. *Am J Hum Genet* **80,** 1024-1036.
17. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100-1104.
18. Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451,** 998-1003.
19. Gerhard P (1986) Historical Geography of New Spain, 1519-1821 (Spanish) (Universidad Nacional Autonoma de Mexico, Mexico City).
20. Gerhard P (1991) La Frontera Sureste de la Nueva España (Universidad Nacional Autónoma de México, Mexico City).
21. Gerhard P (1996) La Frontera Norte de la Nueva España (Universidad Nacional Autónoma de México, Mexico City).
22. Buentello-Malo L, et al. (2003) Genetic structure of seven Mexican indigenous populations based on five polymarker loci. *Am J Hum Biol* **15,** 23-28.
23. Cerda-Flores RM, et al. (1992) Gene diversity and estimation of genetic admixture among Mexican-Americans of Starr County, Texas. *Ann Hum Biol* **19,** 347-360.
24. Cerda-Flores RM, et al. (2002) Genetic admixture in three Mexican Mestizo populations based on D1S80 and HLA-DQA1 loci. *Am J Hum Biol* **14,** 257-263.
25. De Leo C, et al. (1997) HLA class I and class II alleles and haplotypes in Mexican mestizos established from serological typing of 50 families. *Hum Biol* **69,** 809-818.

26. Gorodezky C, et al. (2001) The genetic structure of Mexican Mestizos of different locations: tracking back their origins through MHC genes, blood group systems, and microsatellites. *Hum Immunol* **62,** 979-991.
27. Lisker R, et al. (1986) Gene frequencies and admixture estimates in a Mexico City population. *Am J Phys Anthropol* **71,** 203-207.
28. Lisker R, et al. (1990) Gene frequencies and admixture estimates in four Mexican urban centers. *Hum Biol* **62,** 791-801.
29. Martinez-Marignac VL, et al. (2007) Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. *Hum Genet* **120,** 807-819.
30. Marchini J, et al. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39,** 906-913.
31. Zeggini E, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet.*
32. Patterson N, Price AL, & Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* **2,** e190.
33. Clark AG, et al. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15,** 1496-1502.
34. Falush D, Stephens M, & Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567-1587.
35. Falush D, Stephens M, & Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* **7,** 574-578.
36. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102,** 15942-15947.
37. Rosenberg NA, Li LM, Ward R, & Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73,** 1402-1422.
38. Rangel-Villalobos H, et al. (2008) Genetic admixture, relatedness, and structure patterns among Mexican populations revealed by the Y-chromosome. *Am J Phys Anthropol* **135,** 448-461.
39. Aguirre-Beltran G ed. (1972) La población negra de Mexico. Estudio etnohistórico. (Fondo de Cultura Económica, Ciudad de México).
40. Matsuzaki H, et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1,** 109-111.
41. Conrad DF, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38,** 1251-1260.
42. Cossrow N & Falkner B (2004) Race/ethnic issues in obesity and obesity-related comorbidities. *J Clin Endocrinol Metab* **89,** 2590-2594.
43. Everhart JE, et al. (2002) Prevalence of gallbladder disease in American Indian populations: findings from the Strong Heart Study. *Hepatology* **35,** 1507-1512.
44. Hamman RF, et al. (1989) Methods and prevalence of non-insulin-dependent diabetes mellitus in a biethnic Colorado population. The San Luis Valley Diabetes Study. *Am J Epidemiol* **129,** 295-311.
45. Scheet P & Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78,** 629-644.
46. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81,** 559-575.
47. Bonnen PE, et al. (2006) Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* **38,** 214-217.
48. Barrett JC, Fry B, Maller J, & Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263-265.

Materiales y Métodos Suplementarios

Silva Zolezzi et al. 10.1073/pnas.0903045106

Muestras. Todos los participantes declararon tener cuatro abuelos mexicanos nacidos en el estado del reclutamiento y que no se reconocen como inmigrantes recientes. Todos los Zapotecos son individuos sin relación familiar en primer grado.

Participación comunitaria e individual. Este estudio fue aprobado por los Consejos de Revisión Científica, de Ética y de Bioseguridad del Instituto Nacional de Medicina Genómica (INMEGEN). Se aplicó un proceso ad-hoc para la consulta y participación comunitaria que incluyó tanto a autoridades gubernamentales estatales y autoridades universitarias y sanitarias, como a miembros de la comunidad local. La estrategia se inició dos a tres semanas antes de la obtención de las muestras e incluyó lo siguiente: a) la distribución de un folleto que explica el proyecto escrito en lenguaje sencillo; b) la exhibición de un póster que reproduce el formato de consentimiento informado; c) la organización de cuatro a seis sesiones públicas de carácter informativo, y d) la comunicación por medio de televisión, radio y prensa escrita locales. Todos los participantes otorgaron el consentimiento en presencia de dos testigos locales. En el caso de las muestras de individuos zapotecas, el consentimiento se tradujo a su lengua nativa y todas las partes del proceso se llevaron a cabo en presencia de un traductor bilingüe. En el caso de los mestizos, participaron principalmente, pero no exclusivamente, miembros de la comunidad universitaria estatal local.

Extracción de ADN y genotipificación de genoma completo con SNPs. La extracción de ADN genómico, la genotipificación y el control de calidad de datos se realizaron en el INMEGEN (México). La tasa de genotipificación exitosa promedio utilizando el algoritmo BRLMM, fue 99.45%. En nuestro análisis sólo se incluyeron SNPs presentes en más de 80% de las muestras, con un límite Hardy-Weinberg para valor p de equilibrio mayor a 0.0001, y en los hombres con heterocigocidades en cromosoma X menor al 1%. Los haplotipos ordenados o "phased" para los genotipos del 100K de las muestras HapMap se obtuvieron del HapMap fase II (www.hapmap.org).

Métodos estadísticos. Todos los análisis PCA se realizaron sin ningunas iteraciones eliminando valores atípicos. Las diferencias en contribuciones ancestrales se analizaron con una prueba U de Mann-Whitney. Para examinar si las complejas características estructurales de la población eran responsables de dichas diferencias, se estimó el coeficiente de correlación Pearson entre la FST y las diferencias de la contribución de ancestría promedios (StatPlus:mac versión 2008 de AnalystSoft). Las distribuciones de cajas y bigotes (boxplot) de los componentes ancestrales y los coeficientes de variación, que es una medida normalizada de la dispersión de una distribución de probabilidad, definidos como el índice de la desviación estándar a la media, $CV = \sigma/\mu$, se calcularon utilizando R.

Análisis de componentes ancestrales. Los Marcadores Informativos de ancestría (aims, por sus siglas en inglés) seleccionados, fueron SNPs con diferencias en frecuencia alelica (δ) ≥ 0.4 para cualquier comparación pareada entre grupos HapMap y la población amerindia fuente (CEU, YRI, JPT+CHB, y ZAP). Para minimizar el LD de fondo, se eliminaron marcadores que estuvieran a distancias inferiores a 500Kb de cualquier otro marcador. Estos AIMs dr utilizaron para correr el programa STRUCTURE versión 2.1 utilizando el modelo de ligamiento permitiendo mestizaje, con un número de poblaciones parentales entre K=3 y K=7. Se realizaron diez repeticiones por cada K, con 10,000 ciclos de verificación (burn-in) y 10,000 réplicas sin información poblacional previa (análisis no-supervisado).

Desequilibrio de ligamiento (LD) y Análisis de haplotipos compartidos. La diversidad de haplotipos extendidos (LRHD, por sus siglas en inglés), se estimó diviendo el genoma en ventanas de 35 marcadores en promedio, con una extensión de alrededor de 1 Mb. La diversidad haplotípica se infirió en cada población a través de la comparación del número promedio de haplotipos a lo largo de estas regiones en cada población. Los haplotipos compartidos se analizaron mediante la comparación de frecuencias de haplotipos de 5 SNPs consecutivos cubriendo ~100kb. Aquellos haplotipos que exhibieron una frecuencia mayor a 1% y 5% fueron comparados entre todas las poblaciones analizadas.

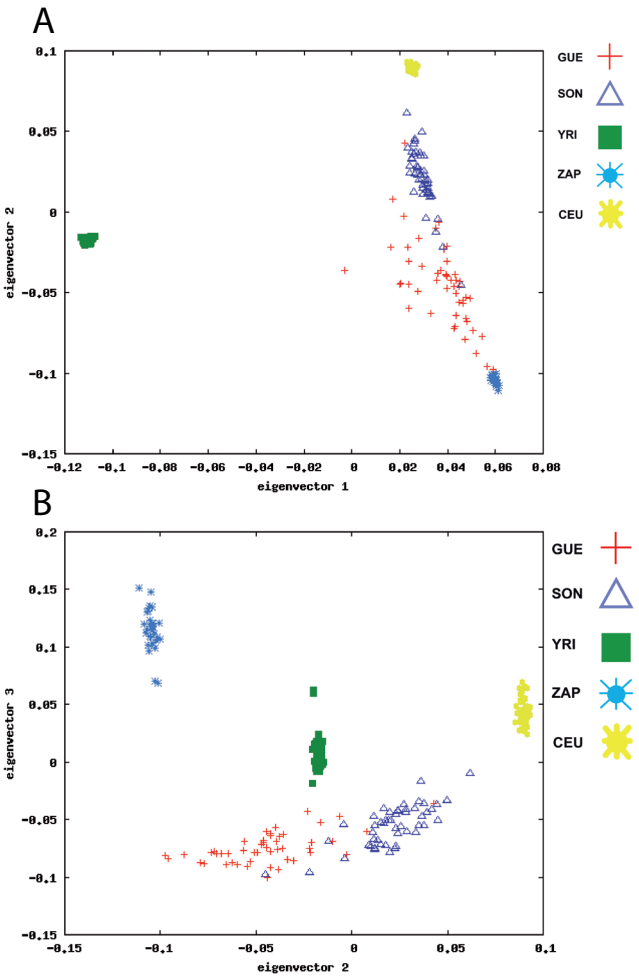

Fig. Suplementaria 1. Análisis de componentes principales. Se graficaron los cuatro Eigen vectores más informativos para el conjunto de datos que incluye subpoblaciones mestizas mexicanas que muestran la mayor diferencia en HET (SON y GUE), ZAP, CEY y YRI. A) Primer y segundo Eigen vectores y B) Segundo y tercer Eigen vectores.

Fig. Suplementaria 2. Distribución de frecuencias de SNPs con el mayor contenido de información entre subpoblaciones mestizas mexicanas. La frecuencia de los 14 SNPs con el contenido de información más elevado se representa en el contexto geográfico de las seis poblaciones mestizas mexicanas analizadas. El color más oscuro indica la región con la frecuencia más alta, y el color más claro, la región con la frecuencia más baja para un marcador particular.

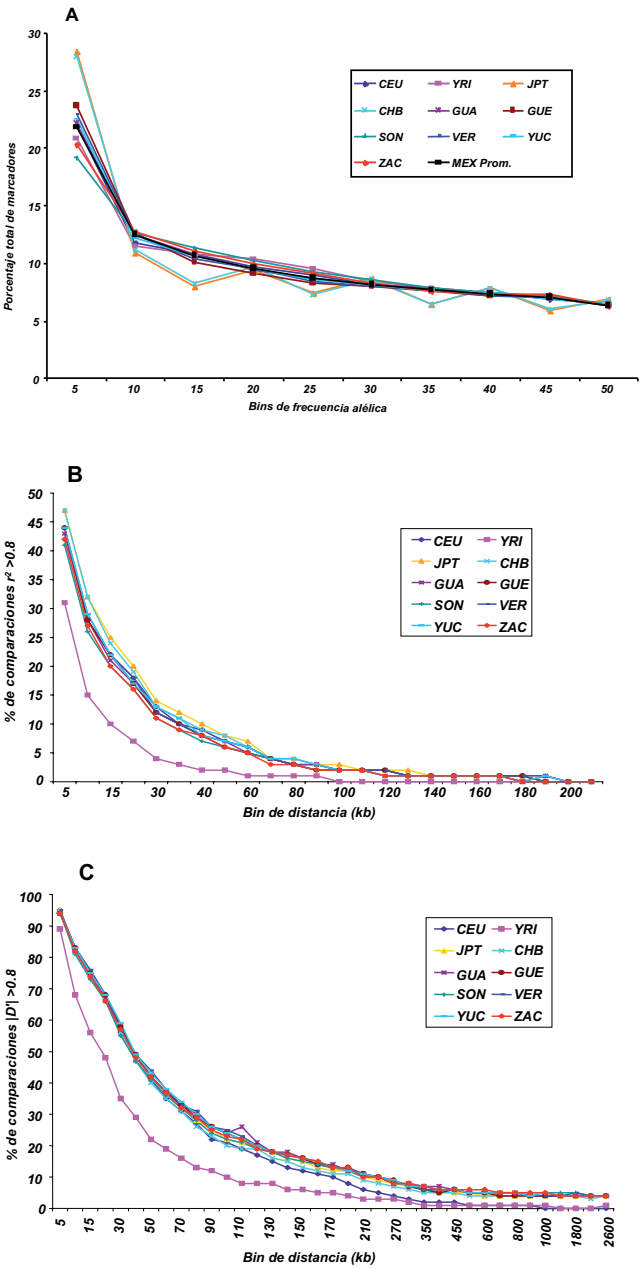

Fig. Suplementaria 3. Distribución de frecuencia de alelos. Distribución de la frecuencia de alelos de los 99,953 SNPs tipificados para las cuatro poblaciones HapMap, las seis mestizas mexicanas e la amerindia.

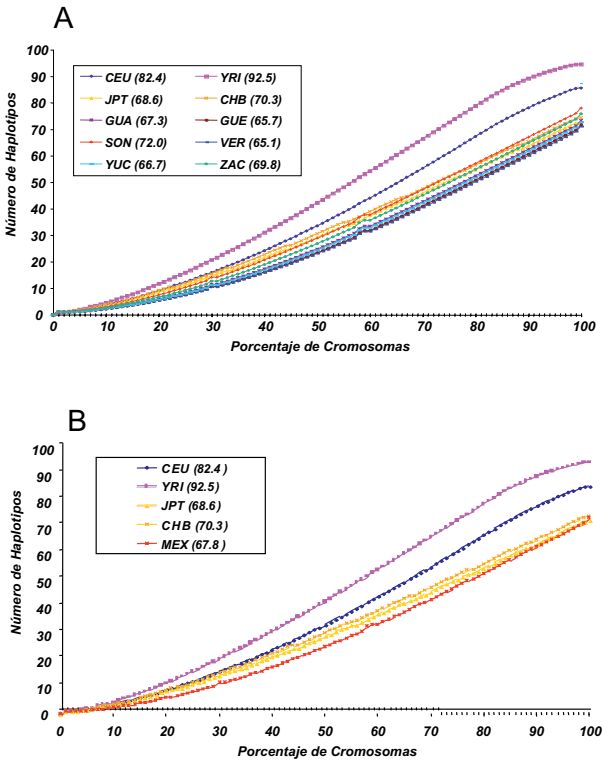

Fig. Suplementaria 4. Decaimiento de LD sobre distancia. Se midió la LD como comparación pareada de marcadores con un MAF≥15% que cayeron en el mismo grupo de distancia. El decaimiento LD sobre la distancia se representa a través del porcentaje de puntajes de comparación pareada iguales a 0.8, para lo que se utiliza A) r2 o B) |D'|.

Tabla Suplementaria 1. Proporciones ancestrales promedio de seis subpoblaciones mestizas mexicanas, amerindios zapoteca de Oaxaca, y todas las poblaciones HapMap.

	STRUCTURE, agrupamiento sin supervisión, $K = 4$, $\delta \geq 0.4$ ($n = 1{,}814$)			
Población	EUR	AMI	AFR	ASI
CEU	0.956 ± 0.024	0.038 ± 0.024	0.002 ± 0.002	0.004 ± 0.005
YRI	0.009 ± 0.007	0.008 ± 0.006	0.981 ± 0.011	0.002 ± 0.003
JPT + CHB	0.017 ± 0.011	0.024 ± 0.020	0.005 ± 0.008	0.954 ± 0.025
ZAP	0.006 ± 0.004	0.992 ± 0.005	0.001 ± 0.001	0.001 ± 0.001
GUA	0.399 ± 0.100	0.576 ± 0.096	0.011 ± 0.018	0.013 ± 0.021
GUE	**0.285 ± 0.120**	**0.660 ± 0.138**	**0.041 ± 0.061**	**0.014 ± 0.021**
SON	**0.616 ± 0.085**	**0.362 ± 0.089**	0.012 ± 0.017	**0.010 ± 0.012**
VER	0.356 ± 0.130	0.613 ± 0.141	0.020 ± 0.042	0.011 ± 0.016
YUC	0.392 ± 0.162	0.588 ± 0.161	**0.008 ± 0.012**	0.012 ± 0.020
ZAC	0.457 ± 0.084	0.511 ± 0.770	0.018 ± 0.023	0.013 ± 0.018
Media mexicana	0.418 ± 0.155	0.552 ± 0.154	0.018 ± 0.035	0.012 ± 0.018

Las proporciones ancestrales se calcularon mediante un conjunto de 1,814 marcadores informativos de ancestría con $\delta \geq 0.4$ después de comparaciones pareadas de cuatro poblaciones parental (CEU, JPT+CHB, YRI, y ZAP). Se resaltan las subpoblaciones mestizas mexicanas que muestran las proporciones más alta y más baja de cada componente ancestral.

Tabla Suplementaria 2. Comparaciones pareadas de estimaciones de ancestría individual para seis subpoblaciones mestizas mexicanas.

	GUA	GUE	SON	VER	YUC
			Ancestría amerindia		
GUE	0.997				
SON	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$			
VER	0.107	0.088	$<1 \times 10^{-4}$		
YUC	0.526	3.1×10^{-2}	$<1 \times 10^{-4}$	0.506	
ZAC	1×10^{-4}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	5.7×10^{-3}
			Ancestría europea		
GUE	$<1 \times 10^{-4}$				
SON	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$			
VER	6.7×10^{-2}	2.50×10^{-3}	$<1 \times 10^{-4}$		
YUC	4.80×10^{-1}	6.00×10^{-4}	$<1 \times 10^{-4}$	3.48×10^{-1}	
ZAC	1.3×10^{-3}	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	$<1 \times 10^{-4}$	1.35×10^{-2}
			Ancestría africana		
GUE	3.83×10^{-1}				
SON	2.44×10^{-1}	9.50×10^{-1}			
VER	3.29×10^{-1}	1.45×10^{-1}	3.04×10^{-2}		
YUC	2.61×10^{-1}	9.12×10^{-2}	2.44×10^{-2}	9.64×10^{-1}	
ZAC	9.06×10^{-2}	8.98×10^{-1}	3.98×10^{-1}	2.74×10^{-2}	5.9×10^{-3}
			Ancestría asiática		
GUE	5.50×10^{-1}				
SON	9.97×10^{-1}	4.65×10^{-1}			
VER	8.20×10^{-1}	3.66×10^{-1}	8.63×10^{-1}		
YUC	5.37×10^{-1}	2.40×10^{-1}	5.06×10^{-1}	9.54×10^{-1}	
ZAC	4.52×10^{-1}	9.42×10^{-1}	3.63×10^{-1}	2.82×10^{-1}	1.74×10^{-1}

Se indican valores P ($p \leq 5x10\text{-}2$) para las diferencias entre las contribuciones de ancestría en cada subpoblación mestiza mexicana.

Tabla Suplementaria 3. Distancia genética entre subpoblaciones mestizas asociadas con diferencias en las principales contribuciones de ancestrales continentales (AMI y EUR).

Pop1-Pop2	Δ_{EUR}	$F_{ST_{Pop1\text{-}Pop2}}$	Distancia genética por ancestría	Distancia genética por ancestría/$F_{ST_{Pop1\text{-}Pop2}}$
GUA-GUE	0.114	2.00*E*-03	2.00*E*-03	1.00
GUE-ZAC	−0.172	5.00*E*-03	4.56*E*-03	0.91
GUE-SON	−0.331	1.90*E*-02	1.69*E*-02	0.89
SON-VER	0.260	1.30*E*-02	1.04*E*-02	0.80
VER-ZAC	−0.101	2.00*E*-03	1.57*E*-03	0.79
GUE-VER	−0.071	1.00*E*-03	7.76*E*-04	0.78
GUA-SON	−0.217	1.10*E*-02	7.25*E*-03	0.66
SON-ZAC	0.159	6.00*E*-03	3.89*E*-03	0.65
SON-YUC	0.224	1.20*E*-02	7.73*E*-03	0.64
GUA-ZAC	−0.058	1.00*E*-03	5.18*E*-04	0.52
GUE-YUC	−0.107	4.00*E*-03	1.76*E*-03	0.44
GUA-VER	0.043	1.00*E*-03	2.85*E*-04	0.28
YUC-ZAC	−0.065	3.00*E*-03	6.51*E*-04	0.22
VER-YUC	−0.036	2.00*E*-03	2.00*E*-04	0.10
GUA-YUC	0.007	3.00*E*-03	7.55*E*-06	0.00

Se estimaron proporciones ancestrales promedio europea (EUR) para seis subpoblaciones mexicanas: (Tabla 1 suplementaria); ΔEUR = proporción promedio de ancestría EUR de Pob1 – proporción ancestral promedio EUR de Pob2: distancia genética entre subpoblaciones = $Fst_{Pop1\text{-}Pop2}$ (Tabla 1); la distancia genética por ancestría se calculó de la siguiente manera: $(\Delta_{EUR})^2 * Fst_{AMI\text{-}EUR}$; $Fst_{AMI\text{-}EUR}$ = 0.154 es la distancia genética entre los amerindios zapotecas (ZAP) y europeos (CEU) del HapMap (Tabla 1).

Tabla Suplementaria 4. Mediana y coeficiente de variación (CV) para estimaciones de componenetes ancestrales individuales en subpoblaciones mestizas mexicanas.

	EUR		AMI		AFR		ASI	
Población	Mediana	CV	Mediana	CV	Mediana	CV	Mediana	CV
GUA	0.378	0.250	0.599	0.167	0.005	1.583	0.005	1.546
GUE	0.289	0.421	0.668	0.210	0.005	1.501	0.006	1.488
SON	0.626	0.139	0.347	0.246	0.005	1.391	0.006	1.264
VER	0.354	0.366	0.608	0.230	0.003	2.096	0.005	1.542
YUC	0.391	0.414	0.595	0.273	0.004	1.527	0.005	1.625
ZAC	0.461	0.184	0.510	0.151	0.009	1.236	0.005	1.322

El CV refleja la varianza normalizada entre individuos de cada grupo. Los valores CV más altos se presentaron para las distribuciones de asiático y africano, lo que indica una varianza mucho mayor en estas contribuciones dentro de cada subpoblación.

Tabla Suplementaria 5. Frecuencia de los 14 SNPs con el mayor contenido informativo (In >0.04) para diferenciar entre grupos mestizos mexicanos, según las estimaciones por informatividad para estadísticas de asignación.

Chr	SNP ID	Posición	Gene	CEU	JHC	YRI	ZAP	GUA	GUE	SON	VER	YUC	ZAC
1	rs4528122	149680414	*POGZ*	0.142	0.856	0.692	0.933	0.650	0.790	0.350	0.690	0.630	0.630
1	rs986690	238518042	*FMN2*	0.250	0.534	0.158	0.983	0.590	0.790	0.360	0.690	0.640	0.650
12	rs6487927	30717602	*IPO8*	0.475	0.466	0.575	0.033	0.270	0.210	0.540	0.130	0.340	0.300
13	rs2147155	92878308	*GPC6*	0.500	0.029	0.025	0.000	0.130	0.092	0.360	0.070	0.080	0.170
4	rs10516422	98483983	—	0.017	0.103	0.183	0.283	0.150	0.350	0.080	0.220	0.337	0.100
5	rs10515716	154822132	—	0.208	0.399	0.133	0.733	0.590	0.650	0.320	0.720	0.470	0.640
6	rs1878071	93533402	—	0.217	0.815	0.008	0.683	0.420	0.670	0.220	0.480	0.480	0.400
9	rs4084051	828022	—	0.175	0.371	0.483	0.750	0.430	0.690	0.270	0.570	0.440	0.420
9	rs7853112	8322353	*PTPRD*	0.350	0.478	0.875	0.750	0.520	0.750	0.310	0.640	0.540	0.510
9	rs10511491	8335923	*PTPRD*	0.392	0.472	0.925	0.750	0.530	0.740	0.320	0.620	0.550	0.520
9	rs1039336	8366287	*PTPRD*	0.242	0.676	0.658	0.867	0.630	0.830	0.430	0.690	0.700	0.510
9	rs10116714	12397578	—	0.050	0.522	0.333	0.817	0.520	0.650	0.229	0.590	0.580	0.460
9	rs1980888	9109037	—	0.100	0.404	0.200	0.966	0.590	0.830	0.360	0.550	0.673	0.470
9	rs4743556	97924766	—	0.167	0.315	0.292	0.828	0.460	0.710	0.306	0.470	0.540	0.450

Tabla Suplementaria 6. Porcentaje de haplotipos comunes compartidos entre las subpoblaciones mexicanas.

Población	GUA GUE	GUA SON	GUA VER	GUA YUC	GUA ZAC	GUE SON	GUE VER	GUE YUC	GUE ZAC	SON VER	SON YUC	SON ZAC	VER YUC	VER ZAC	YUC ZAC
GUA						97	96	96	96	96	97	97	96	96	96
GUE		96	96	96	96					96	96	96	96	96	96
SON	93		93	94	94		92	93	93				93	94	94
VER	96	97		96	97	97		96	96		96	97			97
YUC	95	96	96		96	96	95		96	96		96		96	
ZAC	95	96	95	95		96	95	95		96	96		95		
MEX Prom	95	96	95	95	96	97	95	95	95	96	96	97	95	96	96

Se evaluó el porcentaje de haplotipos compartidos mediante la comparación de frecuencias haplotipos extendidos de 5 SNPs de ~100kb. La tabla muestra el porcentaje de haplotipos comunes compartidos (frecuencia >5%) en donde se utilizan todos los pares posibles de subpoblaciones mexicanas como grupo de referencia contra cada una de las otras subpoblaciones mexicanas.