

Genómica de Poblaciones Colombianas: Ancestralidad y Adaptación.



**Universidad
del Valle**

Augusto Elías Valderrama Aguirre
Universidad del Valle
Escuela de Ciencias Básicas Médicas
Doctorado en Ciencias Biomédicas
Cali, 2016

Genómica de Poblaciones Colombianas: Ancestralidad y Adaptación.

Tesis para optar el título de Doctor en Ciencias Biomédicas

Director:
I. King Jordan, PhD



**Universidad
del Valle**

Augusto Elías Valderrama Aguirre
Universidad del Valle
Escuela de Ciencias Básicas Médicas
Doctorado en Ciencias Biomédicas - 9695
Cali, 2016

Página de aceptación

Jurado 1

Nombre

Firma

Jurado 2

Nombre

Firma

Jurado 3

Nombre

Firma

Cali, _____.

Dedicatoria

A mis hijos David e Isabella...

*A mi madre Hortensia y mis
hermanos Lina y Hernando...*

*A King, amigo y verdadero
consejero...*

*A ese sol que siempre está a mi lado,
ella sabe quién es...*

Contenido

Contenido	6
Listado de Figuras	8
Listado de Tablas.....	9
Resumen.....	10
Capítulo 1	12
Introgresión adaptativa en poblaciones humanas.....	12
Resumen	12
Problema de Investigación.....	12
Pregunta de investigación.....	14
Presentación de la Hipótesis.....	14
Prueba de la Hipótesis y Justificación.....	17
Referencias capítulo 1	19
Capítulo 2	21
Ancestralidad y Mestizaje Genómico en Colombianos.	21
Resumen.....	21
Introducción.....	21
Metodología.....	22
Resultados	25
Discusión.....	28
Referencias Capítulo 2.....	31
Capítulo 3	33
Adaptación Genómica y Salud en Colombianos.....	33
Resumen	33
Introducción.....	33
Metodología.....	34
Resultados	35
Discusión.....	44
Referencias capítulo 3	48
Capítulo 4	51
El Chocó Colombia: un hotspot de biodiversidad humana.....	51
Resumen	51
Introducción.....	51
Metodología.....	53
Resultados	55
Discusión.....	61

Referencias capítulo 4	61
Conclusiones.....	63
Recomendaciones.....	64
Anexos.....	65

Anexo 1: Manuscrito publicado en *Nature Scientific Reviews*.

Anexo 2: Manuscrito publicado en *Revista de Biodiversidad Neotropical*

Anexo 3: Acta de sustentación de tesis doctoral

Listado de Figuras

Capítulo 1	12
Introgresión adaptativa en poblaciones humanas.....	12
Figura 1.1: El intercambio colombino.....	13
Figura 1.2: Modelo para la fijación de alelos adaptativos.....	16
Figura 1.3: Análisis de enriquecimiento ancestral para los eventos de introgresión.....	17
Capítulo 2	21
Ancestralidad y Mestizaje Genómico en Colombianos.	21
Figura 2.1: Descripción de los orígenes geográficos mundiales de los genomas genotipificados y secuenciados alrededor del mundo.....	23
Figura 2.2: PCA de distancias compartidas de alelos apareados entre genomas colombianos de Medellín comparados con poblaciones ancestrales putativas de África, América, Asia y Europa.	26
Figura 2.3: Gráfico de mestizaje genómico que muestra las tres poblaciones ancestrales y la población colombiana mestizada.....	27
Figura 2.4: Distribución de las proporciones de mestizaje ancestral y promedios para la población de Colombia.....	27
Figura 2.5: Contribuciones específicas de sexo al mestizaje genómico triple en varios países de Latinoamérica.....	28
Capítulo 3	33
Adaptación Genómica y Salud en Colombianos.....	33
Figura 3.1: Regiones (locus) y el mestizaje genético ancestral en genomas de Colombia.....	36
Figura 3.2: Distribución de los tamaños de los segmentos cromosómicos con ancestralidad específica usados para el análisis de enriquecimiento.....	36
Figura 3.3: Análisis de enriquecimiento ancestral poblacional.....	37
Figura 3.4: Gráfico de Manhattan mostrando las regiones genómicas identificadas como enriquecidas ancestralmente según la probabilidad trinomial (eje y) de observar una combinación ancestral particular en la población.	38
Figura 3.5: Ejemplo de una región con enriquecimiento ancestral africano hallada en el locus del HLA.....	40
Figura 3.6: Ejemplo de una región con ancestralidad asiática (nativa americana) nula en la que se encuentra el locus del gen EDAR.....	41
Figura 3.7: Análisis de enriquecimiento funcional en regiones con ancestralidad enriquecida.....	43
Capítulo 4	51
El Chocó Colombia: un hotspot de biodiversidad humana.....	51
Figura 4.1: Biodiversidad humana en el Chocó.....	52
Figura 4.2: Ancestralidad y mestizaje genómico en el Chocó.	57
Figura 4.3: Características distintivas del mestizaje en el Chocó.....	58
Figura 4.4: Ancestralidad y mestizaje específico de sexo en El Chocó.....	59
Figura 4.5: Biodiversidad genómica en el Chocó.....	60

Listado de Tablas

Capítulo 2	21
Ancestralidad y Mestizaje Genómico en Colombianos.	21
Tabla 2.1: Perfil demográfico de los grupos étnicos en Colombia y Medellín.	24
Capítulo 3	33
Adaptación Genómica y Salud en Colombianos.....	33
Tabla 3.1: Genes localizados en segmentos cromosómicos enriquecidos,.....	38
identificados en genomas colombianos	38
Tabla 3.2: Genes localizados en regiones enriquecidas ancestralmente y sus características asociadas.	39
Capítulo 4	51
El Chocó Colombia: un hotspot de biodiversidad humana.....	51
Tabla 4.1: Poblaciones analizadas en este estudio.....	56

Resumen

Introducción: El término "intercambio colombino" se refiere a la transferencia masiva de vida entre los hemisferios afro-euro-asiático y americano que fue precipitada por la llegada de Cristóbal Colón al denominado nuevo mundo. Aunque se ha aceptado como una oportunidad para la creación de nuevos genomas humanos, que han sido moldeados por una rápida evolución adaptativa, el efecto de este intercambio en los genomas humanos actuales de los colombianos no ha sido explorado en toda su extensión. El objetivo de esta tesis de doctorado es conocer cuál es el efecto del intercambio colombino en la configuración actual de ancestralidad de los genomas de los colombianos y que implicaciones podría tener en la adaptación al medio de las poblaciones actuales. La hipótesis es que el proceso de evolución del genoma humano, estimulado por el intercambio colombino, se basó en un fenómeno de introgresión selectiva parcial de haplotipos a partir de poblaciones ancestrales, muchos de las cuales tenían utilidades adaptativas pre-evolucionadas.

Metodología: Con base en datos denominados como de alcance genómico (*genome wide*ⁱ), se realizaron análisis comparativos de genomas de poblaciones ancestrales putativas con genomas de poblaciones mestizadas modernas. Se analizaron dos poblaciones humanas de Colombia, una proveniente de la ciudad de Medellín (eurodescendiente) y otra de la ciudad de Quibdó (afrodescendiente). Adicionalmente, se le realizó un análisis de enriquecimiento ancestral a los genomas de Medellín, con el cual se identificaron *loci* con contribuciones ancestrales, a los cuales se les realizó un análisis funcional. Se determinó la variabilidad genética y los haplotipos de genoma uniparentales, mtADN (línea materna) y Y-DNA (línea paterna).

Resultados: Se encontraron haplotipos con ancestralidades específicas que existen en frecuencias más altas en las poblaciones mestizadas de lo que se puede esperar por casualidad, un fenómeno muy similar a la introgresión. Los genomas colombianos de Medellín mostraron que su componente principal de ascendencia es europeo, seguido de nativo americano y africano. Algunos genes y cascadas de reacciones relacionadas con el funcionamiento del sistema inmune adaptativo e innato estaban particularmente sobrerrepresentadas entre los segmentos con ancestralidad enriquecida; así como genes implicados en la adaptación al medio como aquellos relacionados con la pigmentación de la piel (SCL4A5) y glándulas cutáneas (EDAR). La población del Chocó también mostró una mezcla genética triple pero predominantemente africana con contribuciones europeas y nativas americanas. En ambas poblaciones se encontró un exceso de ancestralidad europea en el linaje paterno y, por el contrario, un exceso de nativo americano en el linaje materno.

ⁱ *Genome wide*: término que se usa para indicar que se trata de datos relativos a la secuencia de un genoma completo. Puede tratarse de una secuencia genómica completa o de microarreglos de ADN para caracterizar varios cientos de miles de SNPs alojados en todo el genoma humano.

Conclusiones: El mestizaje genético de los colombianos es el de mayor extensión en los países de Latinoamérica. El mestizaje genético específico de sexo demuestra claramente que el linaje paterno europeo fue dominante en la población analizada. Se demostró que algunos *loci* poseen patrones de ancestralidad específica y que han sido retenidos de forma diferencial en la población colombiana moderna en función de su utilidad en el entorno del llamado Nuevo Mundo. Los resultados sugieren que la población del Chocó posee una profunda diversidad genética humana. Finalmente, los hallazgos permiten subrayar la importancia del papel de la introgresión como fuente de alelos de adaptación y como motor de cambio evolutivo; además, pone de relieve el papel del mestizaje como facilitador de la rápida evolución humana.

Palabras clave:

Genómica en poblaciones colombianas, Introgresión adaptativa, mestizaje genómico, Bioinformática, ChocoGen.

Capítulo 1

Introgresión adaptativa en poblaciones humanas

Resumen

Problema de investigación: El término "intercambio colombino" se refiere a la transferencia masiva de vida entre la los hemisferios Afro-euroasiático y americano que fue precipitada por la llegada de Cristobal Colón al denominado nuevo mundo. El intercambio colombino es muy apreciado por los historiadores, científicos sociales y economistas como un importante punto de inflexión que tuvo efectos profundos y duraderos en la trayectoria de la historia y el desarrollo humano. No obstante, el efecto de este intercambio en los genomas humanos actuales no ha sido explorado en toda su extensión. *Pregunta de investigación:* ¿Cuál podría ser el efecto del intercambio colombino en la configuración actual de ancestralidad de los genomas de los colombianos y que implicaciones podría tener en la adaptación al medio de las poblaciones? *Presentación de la Hipótesis:* Se propone que el intercambio colombino también debe ser apreciado por la comunidad científica por su papel en la creación de nuevos genomas humanos que han sido moldeadas por una rápida evolución adaptativa. En concreto, la hipótesis de trabajo es que el proceso de evolución del genoma humano estimulado por el intercambio colombino se basó en un fenómeno de introgresión selectiva parcial de haplotipos a partir de poblaciones ancestrales, muchos de las cuales tenían utilidades adaptativas pre-evolucionadas, basadas en los efectos sobre la adaptación al medio y la salud en las regiones específicas habitadas por las poblaciones ancestrales. La prueba de esta hipótesis requirió de análisis comparativo de genomas de poblaciones ancestrales putativas con genomas de poblaciones mestizadas modernas. Con esta aproximación se encontraron haplotipos con ancestralidades específicas que existen en frecuencias más altas en las poblaciones mestizadas de lo que se puede esperar por casualidad, un fenómeno muy similar a la introgresión. Un análisis más profundo permitió obtener evidencias sobre el papel funcional de los genes en las zonas con introgresión y de las fuerzas selectivas que han actuado para aumentar su frecuencia en la población. *Conclusión:* Los hallazgos permiten subrayar la importancia del papel de la introgresión como fuente de alelos de adaptación y como motor de cambio evolutivo; además, pone de relieve el papel del mestizaje como facilitador de la rápida evolución humana.

Problema de Investigación

El historiador Alfred Crosby acuñó el término "intercambio colombino" para describir la amplia transferencia de vida entre los hemisferios Afro-euro-asiático (viejo mundo) y América (nuevo mundo) que le siguieron a la llegada de Cristóbal Colón en 1492.¹ El intercambio colombino fue una consecuencia de los esfuerzos de colonización y comercialización europeos en América, lo que requirió una transferencia bidireccional de numerosas especies de plantas, animales y microorganismos entre el viejo y el nuevo mundo (ver Figura 1.1). Esta transferencia también incluyó grupos de poblaciones humanas, culturas y tecnologías, y como tal, dio lugar a grandes cambios demográficos en ambos hemisferios.²

La introducción de especies de cultivos del nuevo mundo, ricas en calorías y nutritivas, como la papa, el maíz y la yuca, en particular, facilitaron desarrollos agrícolas que

permitieron el crecimiento sostenido de la población en el viejo mundo.³ Los cambios demográficos en el nuevo mundo durante este tiempo fueron aún más drásticos. El intercambio colombino reunió a poblaciones previamente aisladas de Europa, África y las Américas en las colonias del nuevo mundo durante un período relativamente corto de tiempo. Más de 50 millones de europeos emigraron a las Américas durante el siglo XIX y el tráfico de esclavos dió lugar a la migración forzada de 12 millones de africanos al nuevo mundo durante un período de ~450 años.^{4,5} La población indígena del nuevo mundo, por el contrario, se redujo hasta en un 95% en un periodo de 100-150 años después del primer viaje de Cristobal Colon, una pérdida que se estima en unas 10-100 millones de vidas;⁶ lo cual se debió en gran medida al resultado de la introducción de enfermedades infecciosas del viejo mundo, como la viruela, el sarampión y la malaria, a las cuales las poblaciones nativas tenían poca o ninguna resistencia. Es por tanto lógico asumir que una profunda transformación demográfica humana podría ocasionar un cambio evolutivo sustancial en el nivel genómico.



Figura 1.1: El intercambio colombino. Ejemplos de plantas, animales y enfermedades infecciosas transferidos entre el viejo y el nuevo mundo durante el intercambio colombino. Las poblaciones humanas de Europa, África y las Américas también se reunieron durante esta época.

Desde la perspectiva de la genómica poblacional, el intercambio colombino se puede considerar como un evento facilitador de la mezcla genéticaⁱⁱ entre los tres grupos de poblaciones humanas –africanos, europeos y nativos americanos– que previamente habían evolucionado por separado durante muchos miles de años.⁷⁻¹⁰ Durante el tiempo en que estas poblaciones estuvieron aisladas, acumularon numerosas diferencias genéticas, en términos de la frecuencia de alelos. Muchas de estas diferencias fueron probablemente cambios neutrales sin efectos apreciables, mientras que otros fueron el resultado de adaptaciones a las presiones de selección locales.^{11,12} En cualquier caso, la

ⁱⁱ A esta mezcla genética, se le seguirá denominando mestizaje genómico para los propósitos de esta tesis de doctorado. El objetivo es mantener la perspectiva humana en el proceso.

acumulación de tales diferencias genéticas resultó en la presencia de haplotipos distintivos; es decir, combinaciones de alelos ligados, que son específicos para las poblaciones individuales. El proceso de mestizaje durante el intercambio colombino llevó luego a la introgresiónⁱⁱⁱ repetida de estos haplotipos específicos de cada población sobre los distintos trasfondos genómicos. En otras palabras, el mestizaje de las poblaciones durante el intercambio colombino generó nuevas secuencias genómicas humanas con combinaciones de haplotipos que nunca habían coexistido previamente en el mismo genoma. Es interesante por tanto explorar las implicaciones de la rápida creación de tales secuencias genómicas, nuevas y mestizadas en el continente americano, en la evolución humana adaptativa.

Aunque la perspectiva de la ancestralidad en Colombianos ha sido previamente abordada en varios trabajos de investigación, esta se ha analizado usando marcadores genéticos que aunque informan sobre las poblaciones ancestrales, lo hacen con una capacidad limitada porque no cubren todo el genoma, ni permiten detectar zonas especialmente enriquecidas con algún trasfondo ancestral particular. El objetivo de esta tesis de doctorado es determinar cuál es el efecto del intercambio colombino en la configuración actual de ancestralidad de los genomas de los colombianos y que implicaciones podría tener en la adaptación al medio de las poblaciones actuales. Es difícil considerar premeditadamente algún efecto puesto que desde 1492 solo han pasado algo más de 520 años, tiempo suficiente para unas 20-25 generaciones (considerando un tiempo promedio de 20-25 años para cada una); por tanto, la hipótesis es que el proceso de evolución del genoma humano, estimulado por el intercambio colombino, se basó en un fenómeno de introgresión selectiva parcial de haplotipos a partir de poblaciones ancestrales, muchos de las cuales tenían utilidades adaptativas pre-evolucionadas.

Pregunta de investigación

¿Cuál podría ser el efecto del intercambio colombino en la configuración actual de ancestralidad de los genomas de los colombianos y que implicaciones podría tener en la adaptación al medio de las poblaciones?

Presentación de la Hipótesis

Introgresión adaptativa y evolución humana rápida: Las mutaciones son la fuente última de nuevos alelos de adaptación y por tanto pueden ser consideradas como un paso limitante de la velocidad crítica de la evolución adaptativa. Las tasas de mutación de la línea germinal humana son bajas y de acuerdo a esto evolución adaptativa en las poblaciones humanas se consideran generalmente un proceso lento que se lleva a cabo durante muchos miles de años.¹²⁻¹⁵ Sin embargo, la introgresión también puede ser una importante fuente de nuevos alelos para la adaptación humana.¹⁶ De hecho, varios estudios recientes han proporcionado evidencia de evolución adaptativa de haplotipos

ⁱⁱⁱ Introgresión: se define como el movimiento de un gen (flujo de genes) de una especie hacia los genes de otra especie por el entrecruzamiento repetitivo con una de las especies parentales.

que sufrieron introgresión a partir de genomas humanos arcaicos (Neandertal y/o Denisova) en los genomas humanos modernos.¹⁷⁻²⁰ La introgresión tiene el potencial de acelerar la evolución adaptativa mediante la introducción de nuevos alelos a una velocidad relativamente rápida en comparación con las mutaciones *de novo*.

Si el mestizaje genético entre poblaciones previamente aisladas es extenso, puede proporcionar haplotipos de introgresión a frecuencias intermedias o altas en la población mestizada resultante. La introgresión podría aumentar así la tasa de evolución adaptativa mediante la elevación de la frecuencia de alelos potencialmente beneficiosos disponibles en la población. En este sentido, la introgresión de alelos de adaptación puede ser considerado como una oportunidad para las denominadas transferencias selectivas suaves (*'soft selective sweeps'*), que han sido definidas como causa de evolución molecular rápida mediante el incremento simultáneo de la frecuencia de múltiples alelos de adaptación en la población.²¹ Las transferencias selectivas suaves pueden ocurrir bajo escenarios evolutivos distintos, incluyendo el caso en que preexisten múltiples alelos de adaptación en la población como variación genética permanente.²² La introgresión en la escala observada para el mestizaje genómico poblacional triple que caracterizó el intercambio colombino podría haber proporcionado múltiples alelos de adaptación como variación genética permanente a frecuencias poblacionales intermedias o altas.⁷⁻¹⁰

Hipótesis de trabajo: el mestizaje genómico y la introgresión entre los grupos de poblaciones humanas que fueron reunidos mediante el intercambio colombino proporcionaron la oportunidad para una rápida evolución adaptativa basada en la existencia de numerosos haplotipos preadaptados. En otras palabras, la introgresión durante el intercambio colombino proporcionó una amplia variación genética permanente a poblaciones del nuevo mundo, muchos de ellos con un potencial significado adaptativo, lo que podría haber proporcionado la materia prima para numerosas transferencias selectivas parciales.

Los tres grupos de poblaciones humanas ancestrales –africanos, europeos y nativos americanos– que se han reunido durante los últimos 500 años durante el curso del intercambio colombino, empezaron a divergir hace ~60-100.000 años, desde que los humanos modernos salieron de África y se difundieron por todo el mundo.²³ Europa fue poblada por seres humanos anatómicamente modernos hace ~40-45.000 años y los seres humanos llegaron a América hace ~15.000 años en varias oleadas de migración a través del estrecho de Bering.^{24,25,26,27} Dado que los tres grupos de población permanecieron aislados durante el curso de la evolución humana, lograron divergir genéticamente, acumulando numerosas diferencias en frecuencias alélicas. Varias de estas diferencias en frecuencias alélicas son probablemente resultado de sustituciones adaptativas con utilidad específica para la región geográfica habitada, en la salud y la adaptación al medio.^{15,11,12} Estos alelos adaptativos preevolucionados, y los haplotipos específicos ancestrales en los que residen, podrían haber sido seleccionado en la población americana

mestizada, en función de su utilidad en el entorno del nuevo mundo. El proceso de selección en este caso se basa en retención diferencial de haplotipos con ancestralidad específica que proporcionen una relativa mejor adaptación en la población mestizada y exclusión de aquellos que no proporcionen ventajas significativas.

El medio ambiente del nuevo mundo que sirvió como crisol selectivo para la introgresión de haplotipos habría consistido tanto del entorno físico externo como del entorno microbiano interna, formado a su vez por la combinación de microorganismos en las poblaciones ancestrales. Particularmente el entorno microbiano nuevo, es probable que haya ejercido una fuerte presión selectiva sobre las poblaciones del nuevo mundo, basado en la necesidad de responder al reto de agentes patógenos infecciosos, lo que sugiere que los genes del sistema inmune podrían ser particularmente propensos a la evolución adaptativa acelerada por la introgresión.²⁶

Cabe señalar que los ~500 años que han transcurrido durante la era del intercambio colombino es una cantidad de tiempo extremadamente corto con respecto a la evolución humana; de hecho, correspondería solo a 20-25 generaciones, asumiendo un tiempo de generación de 20-25 años. Independientemente de la función de la introgresión en el suministro de variación genética adaptativa a una población mestizada, esto no sería suficiente tiempo para permitir la fijación completa de los alelos de adaptación. Por lo tanto, el tipo de evolución adaptativa facilitada por la introgresión aquí propuesto equivaldría a transferencias selectivas parciales (o en curso).²⁷ Sin embargo, niveles sustanciales de cambios en las frecuencias alélicas pueden ocurrir en esta escala de tiempo.

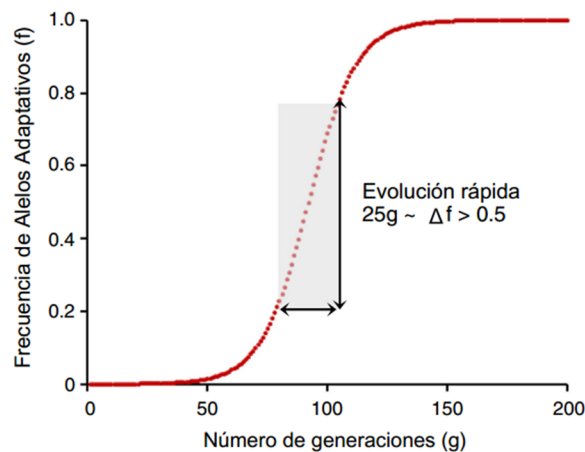


Figura 1.2: Modelo para la fijación de alelos adaptativos. El número de generaciones necesarias para corregir un alelo de adaptación se modela para un coeficiente de selección (s) de 0,01 y un coeficiente de dominancia (h) de 1,0. El nivel de cambio adaptativo por generación en las frecuencias de alelos varía más de cuatro órdenes de magnitud y alcanza su máximo en las frecuencias intermedias de alelos.

Un modelo estándar para la tasa de fijación de un alelo de adaptación ilustra cómo el incremento selectivo en la frecuencia de los alelos procede sucesivamente a través de

regímenes de cambio lento-rápido-lento (ver Figura 1.2). La cantidad de cambio por generación es más alto en las frecuencias intermedias de alelos, en los que la frecuencia de un alelo beneficioso podría aumentar en más de un 50% en 25 generaciones.

Prueba de la Hipótesis y Justificación

Prueba de la hipótesis: La colección de las poblaciones humanas mezcladas de forma única que se encuentra en las Américas representa un laboratorio ideal para estudiar una la rápida adaptación humana y para poner a prueba la hipótesis de la introgresión adaptativa a través del intercambio colombino. Para probar esta hipótesis, sería necesario comparar las secuencias genómicas entre las poblaciones ancestrales putativas con secuencias caracterizadas a partir de poblaciones americanas mestizadas. La comparación con las secuencias genómicas de las poblaciones ancestrales pueden ser utilizadas para caracterizar las contribuciones a las poblaciones mestizadas del nuevo mundo en varios niveles: proporciones de ancestralidad genómicas y específicas de sexo, asignaciones ancestrales locales en los cromosomas y probabilidades de ancestralidad para variantes nucleotídicas individuales (SNPs^{iv}) cuando sea posible. Una vez caracterizadas las proporciones de contribuciones ancestrales globales para cada población americana mestizada, se puede entonces buscar regiones genómicas específicas (loci) o SNPs que se desvíen significativamente de los patrones esperados (ver Figura 1.3).

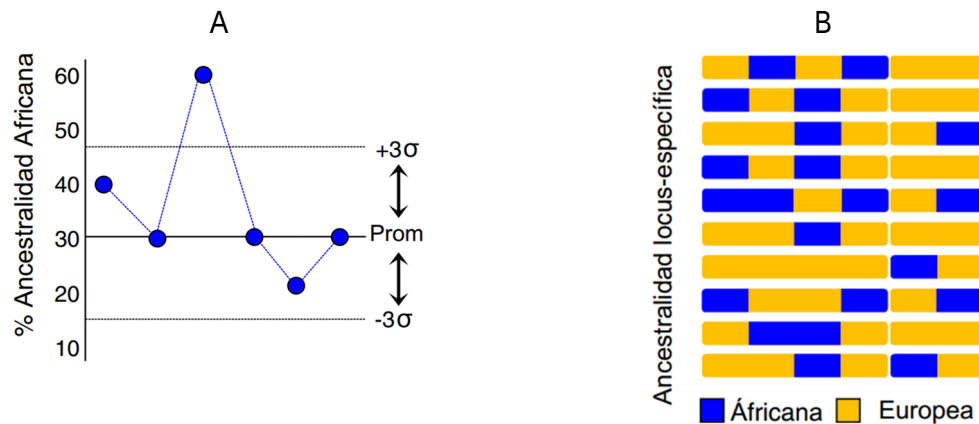


Figura 1.3: Análisis de enriquecimiento ancestral para los eventos de introgresión. A) Se muestra un ejemplo para un único cromosoma de una población hipotética con mestizaje entre dos poblaciones ancestrales: africana (promedio=30%) y europea (promedio=70%). B) Las ancestralidad específica del locus se asigna a todos los cromosomas en la población mestizada y las regiones con niveles anormalmente altos (enriquecidos) o bajas de ancestralidad son identificados para una investigación más detalladas.

Este enfoque puede ser utilizado para identificar aquellos *loci* genómicos que han sido enriquecidos a partir de una población ancestral particular (haplotipos de ancestralidad específica), lo que sugiere la posibilidad de que tales regiones se retuvieron diferencialmente en la población mestizada sobre la base de su utilidad en el medio ambiente del nuevo mundo.^{7,26,28-31} Este enfoque es análogo a la técnica de cartografía

^{iv} SNP: Single Nucleotide Polymorphism

por desequilibrio de mezcla de ligamiento (o mapeo de mezcla), en la que las desviaciones locales de los patrones de mezcla promedio en todo el genoma son usados para identificar *loci* implicados en enfermedades que tienen diferentes prevalencias en poblaciones ancestrales.³² Una vez identificados los *loci* (haplotipos) de ancestralidad enriquecida, estos pueden ser investigados adicionalmente con respecto a sus tasas de evolución, así como las funciones de los genes codificados en ella.

Implicaciones de la hipótesis: La hipótesis de que el intercambio colombino facilitó una rápida evolución adaptativa a través del mestizaje genómico y la introgresión tiene implicaciones tanto para la investigación básica en evolución humana como para investigaciones clínicas sobre los determinantes genéticos de la salud humana. El potencial para la rápida evolución humana es un tema de gran interés y la investigación crítica de la hipótesis propuesta aquí podría ayudar a dilucidar un mecanismo específico por el cual dicha evolución adaptativa rápida puede ser facilitada.^{33,34} Se debe enfatizar que la evolución adaptativa fundamentada sobre retención diferencial de los haplotipos de ancestralidad específica podría implicar cambios bastante sutiles en las frecuencias alélicas a lo largo del rango medio del espectro de frecuencias (ver Figura 1.2). Por lo tanto, se espera que se produzca mucho más rápidamente que la fijación completa de nuevos alelos introducidas por mutaciones *de novo*.

El análisis de las poblaciones mestizadas de América utilizando este marco conceptual tiene el potencial de revelar la evolución humana en acción. Los métodos actuales para detectar las huellas de la evolución adaptativa (ej, Transferencias selectivas) en secuencias del genoma humano se basan en modelos estadísticos complejos de sustitución de secuencias y pueden carecer de poder para distinguir inequívocamente entre los diferentes modelos de selección; por ejemplo las transferencias selectivas suaves vs fuertes, el rol de las variaciones *de novo* frente a las variaciones genéticas permanentes y la prevalencia de selección poligénica.^{22,27,35} En consecuencia, sigue existiendo controversia sustancial en cuanto a la importancia relativa de estos diferentes modos de adaptación en la evolución molecular humana.^{21,35} La alternativa del marco conceptual basado en la introgresión para la detección de haplotipos potencialmente adaptables que se ha mencionado aquí podría ser considerado incierta con respecto a los diferentes modelos del proceso de adaptación, pero también conceptualmente sencilla y concreta, así como sensible a los cambios relativamente menores en las frecuencias de alelos.²⁶

Hasta la fecha, la mayoría de los estudios de genética en las poblaciones humanas del nuevo mundo se han concentrado en la ancestralidad, utilizando variantes de secuencias como marcadores neutrales de los linajes evolutivos. La investigación de la hipótesis que se propone aquí requiere una conexión explícita entre la ancestralidad genética humana y los determinantes genéticos de la salud y la adaptación al medio. La relación entre la ancestralidad y los determinantes genéticos de la salud humana, a menudo se manifiestan como disparidades del perfil salud/enfermedad de las poblaciones,³⁶ es por tanto un tema

importante con consecuencias graves para la salud pública. Por ejemplo, la investigación sobre cómo las poblaciones mestizadas han sido moldeadas por las presiones selectivas impuestas por la carga de enfermedades infecciosas puede dar una idea de la arquitectura genética de la respuesta inmune.³⁷ El estudio de los genomas mestizados a través de todas las Américas, lo cual necesita de la prueba de hipótesis aquí articulada, podría proporcionar una importante extensión a los estudios genómicos clínicos actuales, de los cuales la gran mayoría se han concentrado en las poblaciones de ancestralidad europea.³⁸

Referencias capítulo 1

1. Crosby, A.W. *The Columbian exchange: biological and cultural consequences of 1492*, (Greenwood Publishing Group, Westport, 2003).
2. Mann, C.C. *1493: Uncovering the new world Columbus created*, (Alfred a Knopf Incorporated, New York, 2011).
3. Nunn, N. & Qian, N. The Columbian exchange: A history of disease, food, and ideas. *The Journal of Economic Perspectives*, 163-188 (2010).
4. King, R. *People on the move: An atlas of migration*, (University of California Press, Berkeley, 2010).
5. Segal, R. *The black diaspora: Five centuries of the black experience outside Africa*, (Macmillan, New York City, 1995).
6. Mann, C.C. *1491: New revelations of the Americas before Columbus*, (Alfred a Knopf Incorporated, New York, 2005).
7. Bryc, K., et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 2, 8954-8961 (2010).
8. Montinaro, F., et al. Unravelling the hidden ancestry of American admixed populations. *Nature communications* 6, 6596 (2015).
9. Ruiz-Linares, A., et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS genetics* 10, e1004572 (2014).
10. Wang, S., et al. Geographic patterns of genome admixture in Latin American Mestizos. *PLoS genetics* 4, e1000037 (2008).
11. Grossman, S.R., et al. Identifying recent adaptations in large-scale genomic data. *Cell* 152, 703-713 (2013).
12. Sabeti, P.C., et al. Positive natural selection in the human lineage. *Science* 312, 1614-1620 (2006).
13. Drake, J.W., Charlesworth, B., Charlesworth, D. & Crow, J.F. Rates of spontaneous mutation. *Genetics* 148, 1667-1686 (1998).
14. Boyko, A.R., et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics* 4, e1000083 (2008).
15. Fay, J.C., Wyckoff, G.J. & Wu, C.I. Positive and negative selection on the human genome. *Genetics* 158, 1227-1234 (2001).
16. Hawks, J. & Cochran, G. Dynamics of adaptive introgression from archaic to modern humans. *PaleoAnthropology* 2006, 101-115 (2006).
17. Abi-Rached, L., et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 334, 89-94 (2011).
18. Dannemann, M., Andrés, A.M. & Kelso, J. Adaptive variation in human toll-like receptors is contributed by introgression from both Neandertals and Denisovans. *bioRxiv*.
19. Huerta-Sanchez, E., et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194-197 (2014).
20. Mendez, F.L., Watkins, J.C. & Hammer, M.F. A haplotype at STAT2 Introgressed from neanderthals and serves as a candidate of positive selection in Papua New Guinea. *American journal of human genetics* 91, 265-274 (2012).

21. Messer, P.W. & Petrov, D.A. Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution* 28, 659-669 (2013).
22. Peter, B.M., Huerta-Sanchez, E. & Nielsen, R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS genetics* 8, e1003011 (2012).
23. Meredith, M. *Born in Africa: The Quest for the Origins of Human Life*, (PublicAffairs, New York, 2012).
24. Benazzi, S., et al. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. *Nature* 479, 525-528 (2011).
25. Higham, T., et al. The earliest evidence for anatomically modern humans in northwestern Europe. *Nature* 479, 521-524 (2011).
26. Rishishwar, L., et al. Ancestry, admixture and fitness in Colombian genomes. *Scientific reports* 5, 12376 (2015).
27. Pritchard, J.K., Pickrell, J.K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB* 20, R208-215 (2010).
28. Bhatia, G., et al. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American journal of human genetics* 95, 437-444 (2014).
29. Tang, H., et al. Recent genetic selection in the ancestral admixture of Puerto Ricans. *American journal of human genetics* 81, 626-633 (2007).
30. Jin, W., et al. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome research* 22, 519-527 (2012).
31. Basu, A., et al. Genome-wide distribution of ancestry in Mexican Americans. *Human genetics* 124, 207-214 (2008).
32. Winkler, C.A., Nelson, G.W. & Smith, M.W. Admixture mapping comes of age. *Annual review of genomics and human genetics* 11, 65-89 (2010).
33. Cochran, G. & Harpending, H. *The 10,000 year explosion: How civilization accelerated human evolution*, (Basic Books, New York, 2009).
34. Hawks, J., Wang, E.T., Cochran, G.M., Harpending, H.C. & Moyzis, R.K. Recent acceleration of human adaptive evolution. *Proceedings of the National Academy of Sciences of the United States of America* 104, 20753-20758 (2007).
35. Schrider, D.R., Mendes, F.K., Hahn, M.W. & Kern, A.D. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* 200, 267-284 (2015).
36. Kittles, R.A. & Weiss, K.M. Race, ancestry, and genes: implications for defining disease risk. *Annual review of genomics and human genetics* 4, 33-67 (2003).
37. Karlsson, E.K., Kwiatkowski, D.P. & Sabeti, P.C. Natural selection and infectious disease in human populations. *Nature reviews. Genetics* 15, 379-393 (2014).
38. Rosenberg, N.A., et al. Genome-wide association studies in diverse populations. *Nature reviews. Genetics* 11, 356-366 (2010).
39. Rivas, M.A.M., et al. El Chocó Colombia: a hotspot of human biodiversity. *Revista de Biodiversidad Neotropical* 6, 45-54 (2016).

Capítulo 2

Ancestralidad y Mestizaje Genómico en Colombianos.

Resumen

Introducción: La dimensión humana del intercambio colombino implicó una mezcla genética sustancial entre poblaciones de orígenes ancestrales de África, América y Europa, que habían evolucionado por separado durante miles de años. En esta parte del trabajo se buscó comprender la extensión del mestizaje en los genomas actuales de los colombianos y ancestralidad que estos aún conservan. *Metodología:* Para ello se analizó una población de genomas Americanos mestizados de Medellín, Colombia. A los genomas se les comparó con varias poblaciones ancestrales del planeta y otras nativas americanas mestizadas. La proporción de mestizaje genético específico de sexo también fue explorada. *Resultados:* Los genomas colombianos de Medellín muestran una amplia gama de contribuciones de mezcla de tres orígenes ancestrales. El componente principal de la ascendencia de la población es europea (74,6%), seguido del nativo americano (18,1%) y finalizando en africano (7,3%). Se encontró también que los genomas colombianos reflejan un exceso de ancestralidad europea en el linaje paterno y, por el contrario, exceso de nativo americano en el linaje materno. *Conclusiones:* El mestizaje genético de los colombianos es el de mayor extensión en los países de Latinoamérica con contribución de tres poblaciones ancestrales: europeos, africanos y nativos americanos. Existe una clara divergencia entre la autoidentificación étnica y la ancestralidad genética. El mestizaje genético específico de sexo demuestra claramente que el linaje paterno europeo fue dominante en la población analizada.

Introducción

La llegada de Cristóbal Colón a América precipitó un intercambio masivo y repentino de formas de vida entre los hemisferios Americano y Afro-Eurasia. La colonización y el comercio europeo en América implicaron un intercambio bidireccional de numerosas especies de plantas, animales y microorganismos, muchos de los que hasta ese momento habían evolucionado separadamente durante millones de años. La conquista y colonización de América, junto con el tráfico de esclavos africanos, también incluyó intercambios entre una variedad de poblaciones humanas. Estos intercambios biológicos y culturales, que a menudo eran intencionales aunque a veces no tanto, tuvieron un enorme impacto en la trayectoria de la historia humana en los siglos que siguieron. Entre otros efectos, el intercambio de cultivos y ganado permitió un aumento importante en la población mundial, mientras que la introducción de microbios que causan enfermedades infecciosas tuvo efectos devastadores sobre las poblaciones vírgenes a las que se introdujeron.

El historiador Alfred Crosby se refirió a esta transferencia masiva de vida como el *intercambio colombino*.^{1,2} Desde la introducción del concepto, se han estudiado numerosas dimensiones del intercambio colombino. En particular, los aspectos ecológicos, económicos y culturales e históricos de esta época se han explorado en profundidad. Sin

embargo, un aspecto especialmente intrigante del intercambio colombino que no ha sido abordado dentro de este marco conceptual es el intercambio que se produjo a nivel de secuencias del genoma humano.

La especie humana moderna surgió de África y se extendió por el mundo desde hace 60-100.000 años atrás.³ La evolución humana durante este tiempo se caracterizó principalmente por la migración y el aislamiento geográfico, seguido de divergencia poblacional. A lo largo de decenas de miles de años, estos procesos dieron lugar a los grandes grupos continentales de las poblaciones humanas reconocidas actualmente: África, Europa, Asia, Melanesia y América.^{4,5} Esta trayectoria de la evolución humana se distorsionó durante el intercambio colombino. Durante los últimos 500 años, poblaciones que habían permanecido separadas durante miles de años, fueron agrupadas de nuevo y como resultado comenzaron a intercambiar genes. Este proceso de mezcla genética en América se ha producido durante un periodo de tiempo extremadamente corto en la evolución humana, equivalente a <1% del tiempo transcurrido desde que los humanos surgieron por primera vez en el continente africano. A este proceso de mezcla genética se le denominará **mestizaje genómico**⁵ para los efectos de esta tesis de doctorado.

Esta parte del estudio, se enfocó en determinar el significado de que variantes genómicas, separadas por decenas de miles de años, de repente se vuelven a unir en el transcurso de unos pocos cientos de años. En otras palabras, cuál es el efecto del mestizaje sobre la composición de los genomas modernos de América, particularmente el de los colombianos?

Metodología

Panorama general: Un total de 60 secuencias de genomas de colombianos mestizados de Medellín (Colombia) fueron comparadas con algunas poblaciones de todo el mundo en las que se pudiera asumir una estrecha relación con las poblaciones ancestrales que dieron origen a la población actual de Colombia (ver Figura 2.1). Para ello, se calcularon las distancias de pares alélicos entre SNPs compartidos a través de todos los genomas y se usó un análisis de componentes principales (PCA⁶) para visualizar las distancias apareadas resultantes.

Población y muestra: La población colombiana es particularmente interesante para este tipo de estudios debido a sus altos niveles de mestizaje étnico.⁶ Las personas de Colombia son ricas en diversidad con un mestizaje substancial entre tres poblaciones ancestrales: africanas, nativas americanas y europeas (ver Figura 2.1).⁷⁻¹¹ De hecho, los colombianos han sido reportados previamente como una de las poblaciones con mayor mestizaje continental de estas tres ancestrales, entre las comunidades latino/hispánicas caracterizadas genéticamente.^{12,13} Para efectos de esta parte del trabajo, se analizó la

⁵ https://en.wikipedia.org/wiki/Genetic_admixture

⁶ PCA: Principal component analysis.

ancestralidad y el mestizaje de 60 secuencias genómicas completas de colombianos no relacionados entre sí, originarios de la ciudad de Medellín. Las secuencias de estos genomas, previamente secuenciados con un alto nivel de cubrimiento como parte del proyecto de los 1000 genomas, están disponibles públicamente.^{26,27}

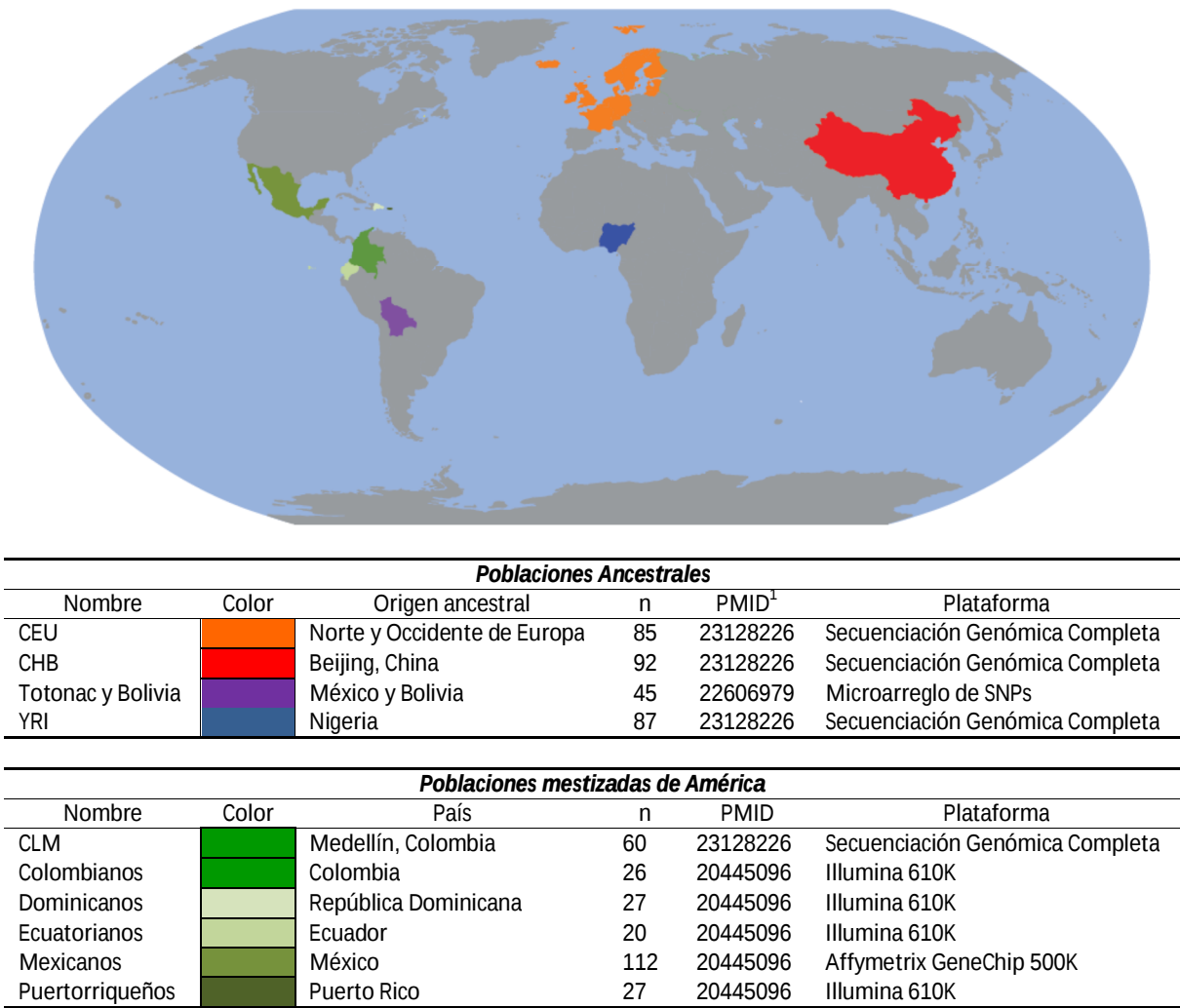


Figura 2.1: Descripción de los orígenes geográficos mundiales de los genomas genotipificados y secuenciados alrededor del mundo. Las localizaciones globales de las poblaciones putativas ancestrales se denotan en colores así: africanos (azul), asiáticos (rojo), europeos (naranja), y nativos americanos (morado). Las poblaciones mestizadas de América se muestran en diferentes tonos de verde. Se indica la información referente a poblaciones originarias, número de muestras, publicaciones y plataforma de caracterización genómica. PMID: *Pubmed Unique Identifier*. Código de identificación único en Pubmed.

Secuencias genómicas: Un total de 581 secuencias genómicas completas fueron extraídas de varias fuentes para ser analizadas en esta parte del trabajo (ver Figura 2.1). Como ya se mencionó, se tomaron 60 secuencias genómicas mestizadas de Colombianos de la ciudad de Medellín a partir de los datos públicamente disponibles del Proyecto de los 1000

genomas.^{26,27} De entre las restantes 521, algunas correspondían a secuencias genómicas completas (n=264) y otras a genotipos caracterizados por microarreglos de SNPs (n=257).¹³⁻¹⁶ Estas secuencias fueron analizadas para establecer sus poblaciones ancestrales putativas en África, América, Asia y Europa. Los genotipos (ej. SNP asignado⁷) caracterizados usando secuencias genómicas y microarreglos a partir de diferentes estudios (y formatos) fueron fusionados usando el programa PLINK v1.90¹⁷, junto con secuencias de comandos personalizadas, con el fin de crear un único conjunto de SNPs fusionados en todos los estudios. A continuación, PLINK se utilizó para realizar una poda de desequilibrio de ligamiento en el grupo de datos de SNP combinados y así producir un conjunto reducido de SNPs no ligados [opciones: `--indep-pairwise 100 25 0.05 --mind 1 --geno 0.01 --hwe .001`]. Estos procesos se hicieron por separado para las secuencias genómicas y los genotipos unidos y para los datos de secuencias genómicas únicamente.

Tabla 2.1: Perfil demográfico de los grupos étnicos en Colombia y Medellín.

Grupo étnico	Colombia (específico) ¹	Colombia (mayoría) ¹	Medellín (mayoría) ¹
Mestizo (amerindio y europeo)	58%	78%	93,4%
Blanco (europeo)	20%	-	-
Mulato (negro/africano y europeo)	14%	21%	6,5%
Negro (africano)	4%	-	-
Zambo (amerindio y negro/africano)	3%	-	-
Amerindio	1%	1%	0,1%

¹Porcentajes de grupos étnicos colombianos, según *CIA World Factbook*. Se detallan valores específicos y de los grupos mayoritarios.

²Porcentajes de grupos étnicos en la ciudad de Medellín, según el censo Colombiano. Solo se muestran los grupos étnicos mayoritarios por ser la única información disponible para las ciudades.

Análisis de ancestralidad y mestizaje: Los alelos que compartieran distancias entre pares de genomas fueron computados como la fracción de las diferencias entre los SNPs asignados. Se realizó un análisis de componentes principales de la matriz de distancias de alelos apareados resultante, usando el programa *prcomp* del paquete R v3.1.2¹⁸ [opciones: `scale=TRUE`] con objeto de relacionar los genomas ancestrales con los mestizados. El programa ADMIXTURE v1.23¹⁹ se utilizó para estimar las fracciones de mestizaje (mezcla) de las tres poblaciones putativas ancestrales – africanos, asiáticos (nativos americanos) y europeos – en los genomas colombianos. El supuesto de que los nativos americanos y los asiáticos son comparables como poblaciones ancestrales es algo que prueba en este mismo trabajo (ver más adelante). El programa ADMIXTURE se utilizó con las opciones que trae por defecto y un k=3 para las poblaciones ancestrales. El programa SupportMix (Ver Jul 18 2012)²⁰ se usó para caracterizar las fracciones de mezclas de las tres poblaciones ancestrales a una escala regional (específico de locus) en los genomas colombianos usando las opciones que trae activadas por defecto. El análisis de mestizaje ancestral específico de locus, se realizó sobre las secuencias genómicas completas obtenidas a partir de los datos disponibles públicamente del proyecto de los 1000 genomas, con objeto de aumentar la resolución (ver Figura 2.1).

⁷ Del Inglés *SNP call*.

Contribuciones al mestizaje específicas de sexo: Relaciones normalizadas de la diferencia entre las fracciones de mezcla para el cromosoma X en comparación con los autosomas fueron utilizadas para deducir las contribuciones al mestizaje específicos de sexo para cada una de las tres poblaciones fuente ancestrales. La relación de diferencias de mestizaje ($\overline{\Delta Admix}$) para cada población ancestral se calculó así:

$$\overline{\Delta Admix} = F_{anc,total} \times (F_{anc,X} - F_{anc,auto}) / (F_{anc,X} + F_{anc,auto}) \quad (1)$$

Donde $F_{anc,total}$ es la proporción de mestizaje de todo el genoma para una ancestralidad dada, $F_{anc,X}$ es la proporción de mestizaje del cromosoma X y $F_{anc,auto}$ es la proporción de mestizaje autosómico para una ancestralidad dada.

Valores positivos de $\overline{\Delta Admix}$ indican un exceso de contribuciones al mestizaje específicamente femeninas (maternas), mientras que valores negativos indican un exceso de contribuciones específicamente masculinas (paternas). Los valores normalizados de los cromosomas X vs los autosomas fueron calculados para los genomas colombianos, así como para datos genómicos previamente reportados de otros 4 países latinoamericanos.¹³ Se utilizó un procedimiento de 1000 réplicas (*bootstrapping*⁸) para crear datos pseudoreplicados con objeto de calcular los intervalos de confianza del 95% para los valores de $\overline{\Delta Admix}$ observados.

Aspectos éticos: Los datos analizados se obtuvieron a partir de bases de datos públicas en las cuales es imposible establecer conexión entre las secuencias genómicas y/o los genotipos y los donantes del ADN humano. Dado que se trata de un trabajo realizado sobre bases de datos públicas, no aplica la solicitud de consentimiento informado, ni la solicitud de aval a un comité de ética en investigación en humanos.

Resultados

Ancestralidad y mestizaje en genomas colombianos: El primer componente principal (81,1% de la variación) muestra una clara separación entre las poblaciones ancestrales africanas y todas las otras poblaciones (ver Figura 2.2), mientras que los nativos americanos, asiáticos y europeos se separan a lo largo del segundo componente (10,5% de la variación). De igual manera, en la figura 2.2 se puede observar que los genomas de Medellín están más relacionados a las poblaciones ancestrales europeas; sin embargo, se separan a lo largo de los dos ejes de los componentes principales: hacia África (PC1) y hacia los nativos americanos (PC2). Esta observación es consistente con el supuesto de la mezcla de ancestralidad africana y nativa americana.

⁸ [https://es.wikipedia.org/wiki/Bootstrapping_\(estad%C3%ADstica\)](https://es.wikipedia.org/wiki/Bootstrapping_(estad%C3%ADstica))

La población asiática está más relacionada con los nativos americanos de México (Totonac) y posteriormente con la población boliviana. Estos últimos comparados a los Totonac, muestran niveles mayores de componente ancestral europeo. La cercanía entre los grupos de población asiática y los nativos americanos (Totonac y Bolivia) es consistente con el origen relativamente reciente de los nativos americanos que llegaron de Asia²¹ y sugiere que los genomas asiáticos son un marcador sustituto viable para inferir ancestralidad nativa americana a nivel continental. En otras palabras, cuando se analizan genomas colombianos mestizados desde una perspectiva de tres orígenes continentales, las comparaciones con poblaciones ancestrales (incluidas las africanas y las europeas) de genomas asiáticos logran recuperar segmentos genómicos ancestrales de nativos americanos. Las secuencias asiáticas (CHB) de genomas completos fueron utilizadas, junto con secuencias genómicas completas de africanos (YRI) y europeos (CEU), para el análisis de genomas colombianos como una forma de proveer una mayor resolución que la disponible a través de los relativamente escasos datos de genotipificaciones.

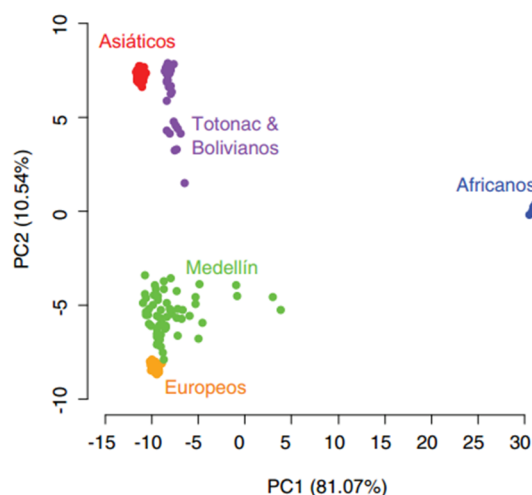


Figura 2.2: PCA de distancias compartidas de alelos apareados entre genomas colombianos de Medellín comparados con poblaciones ancestrales putativas de África, América, Asia y Europa. Los colores indican las poblaciones descritas en la figura 2.1.

Los genomas colombianos muestran una variabilidad substancial en sus patrones de mestizaje con diferentes niveles de contribución de cada una de las poblaciones ancestrales (ver Figura 2.3). El porcentaje de ancestralidad para estos genomas va desde 1,2% africano, 2,1% asiático (nativo americano) y 96,7% europeo hasta 36,8% africano, 18,2% asiático (nativo americano) y 45,0% europeo. Las contribuciones al mestizaje de los genomas colombianos de las poblaciones ancestrales están muy por fuera de los niveles de error observados, esto basado en la extensión de asignación de ancestralidad falsamente positiva (ver Figura 2.3). El promedio de valores de ancestralidad para los genomas mestizados de Medellín, Colombia son 7,3% africano, 18,1% asiático (nativo

americano) y 74,6% europeo, lo que confirma que esta población tiene una ancestralidad principalmente de origen europeo (Figura 2.4).

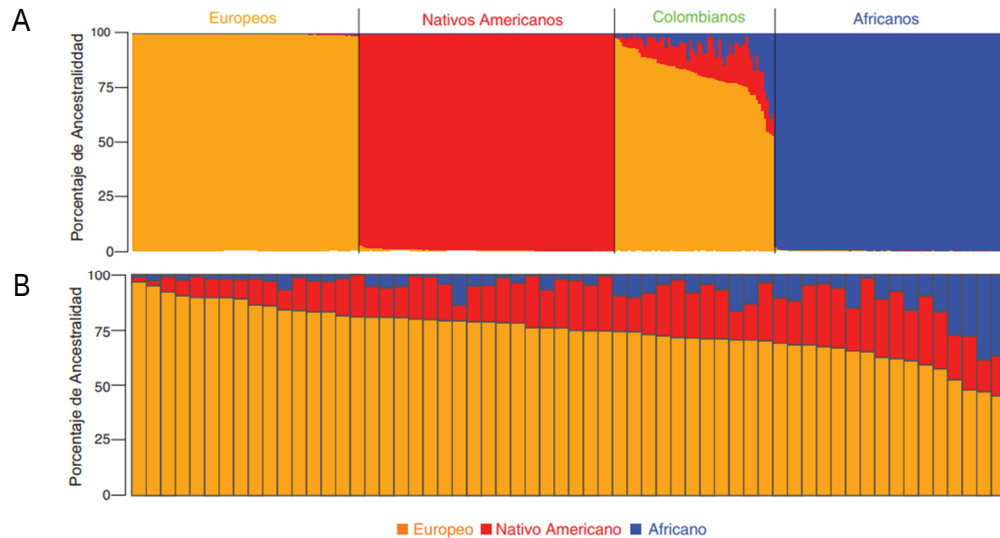


Figura 2.3: Gráfico de mestizaje genómico que muestra las tres poblaciones ancestrales y la población colombiana mestizada. (A). En el panel B se muestra la población de genomas colombianos mestizados solamente.

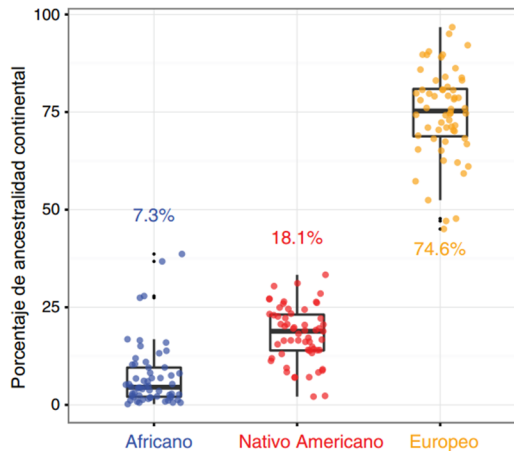


Figura 2.4: Distribución de las proporciones de mestizaje ancestral y promedios para la población de Colombia.

Mestizaje específico dependiente del sexo: varios estudios previos han documentado una asimetría dependiente del sexo en las contribuciones de las poblaciones ancestrales al mestizaje de los genomas en Latinoamérica.^{8-11,13}. Los genomas mestizados tienden a mostrar un relativo exceso de ancestralidad paterna europea y proporcionalmente una mayor ancestralidad materna de origen nativo americano. En este orden de ideas, se evaluó la asimetría sexual del mestizaje en la población Colombiana evaluada comparando las contribuciones ancestrales al cromosoma X vs los autosomas. Dado que el cromosoma X permanece relativamente más tiempo junto al linaje femenino, un exceso relativo específico de algún componente ancestral en el cromosoma X indica una contribución

proporcionalmente mayor de origen femenino (materno) para ese ancestro. Por el contrario, un exceso relativo de mezcla ancestral específica en los autosomas indica una mayor contribución masculina (de origen paterno). De forma consistente con los reportes previos, la población Colombiana muestra unos altos patrones de mestizaje específico de sexo, con contribuciones predominantemente europeas al linaje masculino y predominantemente nativa americana al linaje femenino (ver Figura 2.5).

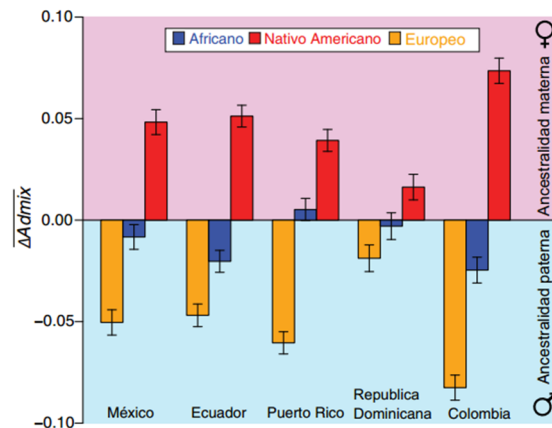


Figura 2.5: Contribuciones específicas de sexo al mestizaje genómico triple en varios países de Latinoamérica. Para cada componente ancestral – africano, asiática (nativo americano) y europeo – se muestra la diferencia normalizada entre la fracción ancestral del cromosoma X y la fracción ancestral autosómica. Los valores positivos en la figura indican un exceso relativo de ancestralidad femenina (materna) para cada componente dado de la mezcla, mientras que los valores negativos indican un exceso de ancestralidad masculina (paterna).

De forma llamativa, este patrón es más pronunciado para la población colombiana que para cualquiera de las otras cuatro poblaciones latinas con las que se comparó. El nivel relativo de ancestralidad masculina europea para la población colombiana es significativamente mayor que la observada para la del país con el siguiente nivel más alto (Puerto Rico; $P=1.8 \times 10^{-7}$). Del mismo modo, el nivel relativo de ancestralidad femenina nativa americana fue significativamente mayor para la población colombiana en comparación con el país con el siguiente nivel más alto (Ecuador; $P=1.2 \times 10^{-7}$). La población de Colombia también muestra un exceso de contribución ancestral masculina de origen africano, menor a la europea pero no despreciable y superior a la de otros países latinos comparándose solo con Ecuador (la ascendencia paterna africana difirió de 0 con $P=3.2 \times 10^{-14}$).

Discusión

Ancestralidad genética en Colombia: Los resultados de este análisis de la secuencia de 60 genomas colombianos de Medellín señala altos niveles de mestizaje genómico, a nivel continental, de tres orígenes; con contribuciones de poblaciones ancestrales africanas, nativas americanas y europeas, consistente con varios estudios genéticos previos y con las tendencias demográficas generales en Colombia (ver Tabla 2.1).^{7-11,13} En promedio, los

genomas mestizados de Medellín mostraron una ancestralidad predominantemente europea; las secuencias genómicas muestran en promedio 74,6% de origen europeo, un 18,1% de Asia (nativo americano) y el 7,3% africano (ver Figura 2.4). Sin embargo, los individuos de Medellín varían ampliamente con respecto a sus proporciones de estas tres poblaciones ancestrales. Hay individuos con >95% de ancestralidad europea en un extremo del espectro y personas con contribuciones mayores de las tres poblaciones ancestrales en el otro extremo (ver Figura 2.3B).

Cabe señalar que en Colombia las diferentes regiones muestran patrones demográficos muy distintos entre sí. Por ejemplo, mientras que la gente de Medellín tiene una ancestralidad principalmente europea, las regiones costeras del Atlántico y del Pacífico son el hogar de poblaciones mayoritariamente afrocolombianas. Por lo tanto, los patrones de mestizaje genómico que se reportan aquí para la ciudad de Medellín no son representativos de los patrones de ancestralidad colombiana en todo el país. De hecho, los estudios sobre la ancestralidad genética de los colombianos con muestras de diferentes regiones del país arrojan a menudo muy diferentes resultados^{19-23,25}. Un manuscrito recientemente publicado podría ser el mayor estudio sobre ancestralidad genética en Latinoamérica hasta la fecha.¹² En ese estudio, el mestizaje de tres orígenes se infiere a partir de 1.659 individuos con base en una colección de 30 marcadores genéticos de ancestralidad (a diferencia de este trabajo en el que se usaron más de 16 millones de SNPs para definir la ancestralidad y el mestizaje). Los resultados del estudio ponen de relieve la diversidad genética de la población de Colombia. Las muestras colombianas mostraron los niveles más altos de mestizaje genómico triple a partir de las poblaciones ancestrales (60% europeos, el 29% de nativos americanos y el 11% de África) entre los cinco países latinoamericanos evaluados, así como el mayor grado de variación geográfica en la ancestralidad genética.

Ancestralidad e identidad en una población colombiana: Los resultados de ancestralidad y mestizaje genómico obtenidos en este trabajo para la población de Medellín pueden considerarse en contradicción con los datos demográficos de la ciudad. En el censo de 2005, el 93,4% de la población de Medellín fue clasificada como eurodescendiente, mientras que el 6,5% de los individuos fueron identificados como afrocolombianos y sólo el 0,1% identificado como Nativo Americano.²² Estos datos demográficos se basan en la autoidentificación y reflejan los grupos étnicos en los que individuos se consideran a sí mismos como miembros. Por lo tanto, parece que la gran mayoría de los colombianos de Medellín se identifican como blancos a pesar de la presencia de una fracción sustancial de los individuos con niveles apreciables de ascendencia nativa americana y africana. Por ejemplo, 46 (~77%) de los individuos colombianos estudiados aquí tienen >2% de ascendencia africana. Dado que el 93,4% de la población de Medellín se autoidentifica como blanco, es probable que la mayoría de los individuos estudiados aquí (~43) se consideren a sí mismos como blancos. Si asumimos que estos 43 individuos ocupan el extremo inferior de la distribución de la ancestralidad africana, entonces el 72% de los

colombianos autoidentificados como blancos de Medellín tienen >2% de ancestralidad africana. A modo de comparación, un reciente análisis a gran escala de ancestralidad genética entre los diferentes grupos étnicos en los Estados Unidos mostró que sólo el 1,4% de los autoidentificados de origen europeo tiene al menos un 2% de ancestralidad africana.²³ También hay que señalar que los individuos colombianos estudiados aquí tienen niveles sustancialmente más altos de nativos americanos en comparación con la ancestralidad africana. Sin embargo, no es posible llevar a cabo este mismo tipo de comparación entre la autoidentificación étnica y la ancestralidad genética con respecto a la ascendencia nativa americana dada la categorización de datos étnicos a nivel de ciudad, según el censo de Colombia.²² Esto se debe a que el censo no distingue entre los descendientes de europeos que se identifican como blancos en comparación con los individuos con una combinación de ancestralidad europea y nativa americana que se identifican como mestizos a nivel de las ciudades individuales.

La diferencia entre ancestralidad genética y autoidentificación étnica encontrada en Medellín puede estar relacionada a dos conceptos culturales importantes, arraigados de manera muy profunda en muchas sociedades latinoamericanas: Mestizaje y Blanqueamiento.²⁴ El “mestizaje”, hace referencia a una mezcla de diferentes grupos étnicos y como tal se considera como una parte crítica de la construcción de la nación y de la identidad cultural para Colombia y muchos otros países de Latinoamérica.^{25,26} Por otra parte, el “blanqueamiento” se refiere a la ideología del mejoramiento racial a través de un aclaramiento de la piel de la población. Aunque el blanqueamiento puede tener una dimensión biológica, con respecto al deseo de tener una descendencia de piel más clara, este se manifiesta como un constructo social comúnmente. Es en este sentido social que el blanqueamiento se refleja como una autoidentificación étnica. Si la blancura se mantiene de manera implícita como un ideal social, y una tendencia progresiva generacional a la que una sociedad debe aspirar, las personas tenderán a autoidentificarse como blancos, independientemente de su ancestralidad genética.²⁷ El contraste entre los resultados obtenidos de ancestralidad genética y los datos demográficos, ejemplifica la posibilidad de que en la población de Medellín este fenómeno sea una tendencia real. Por otra parte, dado que la gran mayoría de los individuos estudiados aquí (95%) tienen genomas con mayor ascendencia europea, estos individuos pueden elegir simplemente identificarse como de ascendencia blanca o europea a pesar de su mezcla genética.

Las relaciones entre ancestralidad genética y autoidentificación étnica han sido recientemente estudiadas en profundidad para 5 países latinoamericanos.¹² Dicho trabajo confirmó que existen grandes variaciones en la ancestralidad genética entre los grupos autoidentificados étnicamente en todos los países estudiados y revela la gran influencia que tiene la apariencia física en la autoidentificación. También demostró que particularmente la pigmentación de la piel tiene un efecto profundo en la autoidentificación étnica, pero también que los rasgos físicos y la apariencia son indicadores pobres de la ancestralidad genética. No obstante, la relación entre

ancestralidad genética y autoidentificación ha demostrado ser compleja. Existen correlaciones consistentes, aunque débiles, entre la ancestralidad genética y los rasgos físicos en latinoamericanos, y la profundidad con la cual la ancestralidad genética es sobre o sub estimada varía entre los grupos étnicos de acuerdo a la categoría étnica y el componente ancestral particular (africano, europeo o nativo americano). Llamativamente, la autopercepción de ancestralidad europea tiende a sobreestimar la ancestralidad europea determinada, mientras que la autopercepción de ancestralidad africana subestima la determinada. La autopercepción de nativo americano subestima el alcance de la ancestralidad genética a niveles más bajos de ancestralidad genética nativa americana, comparable a lo observado en este estudio.

Mestizaje genético y la conquista de América: Las secuencias de los genomas colombianos analizados aquí muestran patrones asimétricos específicos de sexo en el mestizaje genético con linajes masculinos dominados por la ancestralidad europea y linajes femeninos compuestos por ancestralidad principalmente nativa americana (ver Figura 2.5). Estos resultados son consistentes con un número previo de estudios que demuestran patrones similares de mestizaje específico de sexo en colombianos.^{8-11,13} Sin embargo, en este trabajo se demuestra por primera vez que el patrón de ancestralidad específica de sexo es más pronunciado para Colombia que para otros países latinoamericanos. Este patrón de ancestralidad genética refleja la difícil realidad del periodo de “la conquista” y el periodo colonial que siguió. Los conquistadores españoles eran guerreros profesionales que llegaron a América sin sus familias (esposas), y los grupos subsecuentes de colonos europeos incluyeron números relativamente escasos de mujeres.⁹ Adicionalmente, los conflictos armados entre conquistadores y grupos indígenas resultaron en la determinación de una gran parte de la población nativa americana, en la cual los hombres eran un blanco particular de eliminación.²⁸ Estos factores demográficos e históricos resultaron en que los hombres europeos frecuentemente tenían hijos con mujeres indígenas, lo cual justifica la tendencia de mestizaje genético específico de sexo que se observó.

Referencias Capítulo 2

1. Crosby, A. *The Columbian Exchange: Biological Consequences of 1492*, (Greenwood, Westport, 1972).
2. Mann, C.C. *1493: Uncovering the new world Columbus created*, (Alfred a Knopf, New York, 2013).
3. Garrigan, D. & Hammer, M.F. Reconstructing human origins in the genomic era. *Nature reviews. Genetics* 7, 669-680 (2006).
4. Li, J.Z., *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104 (2008).
5. Rosenberg, N.A., *et al.* Genetic structure of human populations. *Science* 298, 2381-2385 (2002).
6. CIA. *CIA World Factbook: Colombia*. (2014).
7. Wang, S., *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS genetics* 4, e1000037 (2008).

8. Cordoba, L., *et al.* Composicion genetica de una poblacion del suroccidente de Colombia. *Revista Colombiana de Antropologia* 48, 21-48 (2012).
9. Carvajal-Carmona, L.G., *et al.* Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *American journal of human genetics* 67, 1287-1295 (2000).
10. Carvajal-Carmona, L.G., *et al.* Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Human genetics* 112, 534-541 (2003).
11. Bedoya, G., *et al.* Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7234-7239 (2006).
12. Ruiz-Linares, A., *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS genetics* 10, e1004572 (2014).
13. Bryc, K., *et al.* Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 2, 8954-8961 (2010).
14. Genomes Project, C., *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073 (2010).
15. Genomes Project, C., *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65 (2012).
16. Watkins, W.S., *et al.* Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC genetics* 13, 39 (2012).
17. Purcell, S., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81, 559-575 (2007).
18. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* 5, 299-314 (1996).
19. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19, 1655-1664 (2009).
20. Omberg, L., *et al.* Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC genetics* 13, 49 (2012).
21. Reich, D., *et al.* Reconstructing Native American population history. *Nature* 488, 370-374 (2012).
22. DANE. Censo General 2005: Perfil Medellin Antioquia. (2005).
23. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D. & Mountain, J.L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American journal of human genetics* 96, 37-53 (2015).
24. Wade, P. *Blackness and race mixture: the dynamics of racial identity in Colombia*, (JHU Press, Baltimore, 1995).
25. Chavez, M. & Zambrano, M. (eds.). *From blanqueamiento to reindigenizacion: Paradoxes of mestizaje and multiculturalism in contemporary Colombia*, 5-23 (2006).
26. Andrews, G.R. *Afro-Latin America 1800-2000*, (Oxford University Press, New York, 2004).
27. Telles, E. & Flores, R. Not just color: Whiteness, nation, and status in Latin America. *Hispanic American Historical Review* 93, 411-449 (2013).
28. Markham, C. *The Conquest of New Granada*, (EP Dutton and Company, New York, 1912).

Capítulo 3

Adaptación Genómica y Salud en Colombianos.

Resumen

Introducción: El intercambio colombino implicó la mezcla genética sustancial entre poblaciones ancestrales de África, América y Europa. Conociendo ya que la población de genomas analizados de la ciudad de Medellín, tienen un componente ancestral principalmente europeo, pero con sustanciales aportes de africanos y nativos americanos, en esta segunda parte del trabajo se exploraron las posibles consecuencias de la creación de genomas americanos mestizados, con nuevas combinaciones alélicas, en la salud humana y la adaptación al medio. *Metodología:* Para ello, a los 60 genomas mestizados Medellín (Colombia), se les realizó un análisis de patrones específicos de ancestralidad a nivel de locus con el objetivo de identificar regiones genómicas enriquecidas en toda la población con contribuciones ancestrales particulares. *Resultados:* Se encontró que especialmente algunos genes y cascadas de reacciones relacionadas con el funcionamiento del sistema inmune adaptativo e innato estaban particularmente sobrerrepresentadas entre los segmentos con ancestralidad enriquecida, incluyendo genes como *HLA-B* y *MAPK10* que están involucrados en la defensa contra patógenos tropicales endémicos como la malaria. Otros genes implicados en la adaptación al medio como aquellos genes que codifican funciones relacionadas con la pigmentación de la piel (*SCL4A5*) y las glándulas cutáneas (*EDAR*) también se encuentran en las regiones con patrones de ancestralidad anormales. *Conclusión:* Estos resultados sugieren la posibilidad de que algunos *loci* de ancestralidad específica han sido retenidos de forma diferencial en la población colombiana moderna, genómicamente mezclada, en función de su utilidad en el entorno del llamado Nuevo Mundo.

Introducción

El mestizaje genómico ocurrido durante el intercambio colombino logró mezclar rápidamente alelos poblacionales específicos (enriquecidos) que no habían coexistido previamente en el mismo trasfondo genético y por lo tanto un resultado que se puede considerar es la creación de genomas humanos completamente nuevos. Los alelos específicos (enriquecidos) de las poblaciones humanas son variantes de secuencias que evolucionaron *in situ* hasta las frecuencias características de dichas poblaciones, de manera restringida a las regiones geográficas donde las poblaciones ancestrales eran endémicas. Algunos de estos alelos pueden haberse fijado como alelos de alta frecuencia dentro de las poblaciones ancestrales por casualidad, mientras que otros son propensos a haber sido eliminados de la fijación con base en las presiones de selección que fueron diferentes en los ambientes específicos de las poblaciones ancestrales.¹⁻⁴ Los genes que median en la interacción de los seres humanos con su medio ambiente, tales como los que codifican la pigmentación de la piel y de las proteínas relacionadas con el sistema inmune, parecen haber sido particularmente propensos a la evolución adaptativa en las regiones ancestrales. Por ejemplo, hay varios genes de pigmentación de la piel con alelos específicos de Europa y/o a Asia que se asocian con tonos de piel claros.^{1-3,5-11} Por otro lado, un número de genes implicados en la defensa contra agentes patógenos infecciosos

también evolucionaron con alelos específicos de la región geográfica, concordantes con los rangos de patógenos particulares de dicha región.¹⁻⁵ Por ejemplo, se han descrito al menos tres genes distintos que han evolucionado alelos específicos poblacionales relacionados con la defensa inmune a la malaria en regiones endémicas para esta enfermedad.^{2,3,5} En general, las poblaciones de las regiones del mundo ricas en patógenos, incluidas regiones de África occidental, Asia oriental y América, codifican un repertorio más diverso de receptores del sistema inmune y esto se piensa que es debido a la presión de selección para hacer frente a una variedad más amplia de patógenos microbianos.¹²

En esta parte del trabajo se explora como afecta a la salud humana y la adaptación al medio el hecho de mezclar, sobre el curso de unos pocos cientos de años, variantes genómicas que han permanecido separadas por decenas de miles de años? En otras palabras, está relacionado el proceso de mestizaje genómico moderno con la salud humana y la adaptación al medio? La hipótesis de trabajo en este caso está basada en el suficientemente comprobado axioma de que las variantes alelo específicas (por ejemplo SNPs⁹) evolucionaron separadamente en las poblaciones humanas ancestrales con base en su utilidad dentro de una región geográfica específica, es decir su relación con la salud y la adaptación a un medio ambiente particular. Aquí se propone que tales alelos específicos de poblaciones ancestrales pueden haber sido sujeto de selección en las poblaciones mestizadas modernas basado en su utilidad en un nuevo medio ambiente. Tal nuevo medio ambiente puede ser un nuevo medio ambiente físico, para poblaciones transportadas a regiones geográficas completamente nuevas y/o un nuevo ambiente ecológico basado en la mezcla de poblaciones humanas previamente aisladas y su fauna microbiana asociada.

Metodología

Panorama general: En esta parte del trabajo se desarrolló y se aplicó un método para buscar regiones genómicas en la que se evidenciaran patrones anormales de mestizaje genómico basado en la retención diferencial de segmentos cromosómicos de poblaciones ancestrales específicas en las poblaciones mestizadas modernas. Estas regiones con patrones anormales fueron exploradas para detectar tanto huellas de selección natural como los roles funcional y asociados a la salud humana de los genes codificados en ellas.

Población y muestra: Se analizó la ancestralidad y el mestizaje de 60 secuencias genómicas completas de colombianos no relacionados entre sí, originarios de la ciudad de Medellín. Las secuencias de estos genomas, previamente secuenciados con un alto nivel de cubrimiento como parte del proyecto de los 1000 genomas, están disponibles públicamente.^{26,27}

⁹ SNP: Single Nucleotide Polymorphism.

Análisis de enriquecimiento ancestral: Las regiones cromosómicas con patrones anormales de ancestralidad, cuando se consideran como un conjunto a través de toda la población de 60 genomas colombianos (120 cromosomas), fueron identificadas a través de valores de probabilidad de enriquecimiento ancestral. Estos valores fueron calculados como una probabilidad trinomial de observar una combinación numérica particular de segmentos cromosómicos africanos (YRI), asiáticos (CHB) y europeos (CEU) para toda la población en un locus genómico específico, dados los niveles promedio de ancestralidad africana, asiática (nativos americanos) y europea. Dicha probabilidad se expresa como:

$$P(YRI = x_1 \text{ and } CHB = x_2 \text{ and } CEU = x_3) = \frac{120!}{x_1! \times x_2! \times x_3!} p_{YRI}^{x_1} \times p_{CHB}^{x_2} \times p_{CEU}^{x_3}$$

Donde x_1, x_2, x_3 son los recuentos de segmentos cromosómicos observados en un locus en particular con ancestralidad YRI, CHB y CEU, y $p_{YRI}, p_{CHB}, p_{CEU}$ son las probabilidades esperadas según los valores de ancestralidad promedio de la población.

Las regiones genómicas con un patrón anormal de ancestralidad significativamente alto ($P < 10^{-9}$, FDR q-value 6.5×10^{-9}) fueron exploradas para conocer su potencial funcional usando varias fuentes de información. Los genes que estaban dentro de estas regiones fueron comprados con grupos de genes previamente identificados en los que se conocía que habían sido sujeto de selección positiva (adaptativa) en varios estudios^{1-5,13} y con genes implicados en procesos de salud y enfermedad humana, a través de estudios de asociación usando revisión de literatura y el catálogo de estudios GWAS¹⁰ (Mayo de 2014) del NHGRI^{11,14}. Las afinidades funcionales de los genes codificados en estas regiones fueron también evaluadas con análisis de enriquecimiento de grupos de genes usando el software GSEA¹² v3.87 vía web.¹⁵

Aspectos éticos: Los datos analizados se obtuvieron a partir de bases de datos públicas en las cuales es imposible establecer conexión entre las secuencias genómicas y/o los genotipos y los donantes del ADN humano. Dado que se trata de un trabajo realizado sobre bases de datos públicas, no aplica la solicitud de consentimiento informado, ni la solicitud de aval a un comité de ética en investigación en humanos.

Resultados

Enriquecimiento del mestizaje genómico a nivel poblacional: Se determinaron los tres orígenes ancestrales continentales de segmentos cromosómicos individuales, a través de todo el genoma, para los 60 individuos en la población colombiana estudiada. En la figura 3.1 se muestran dos ilustraciones de cromosomas coloreados que muestran la localización de los segmentos con ancestralidad específica en dos individuos colombianos.

¹⁰ GWAS: Genome Wide Association Studies.

¹¹ NHGRI: National Human Genome Research Institute.

¹² GSEA: Gene Set Enrichment Analysis.

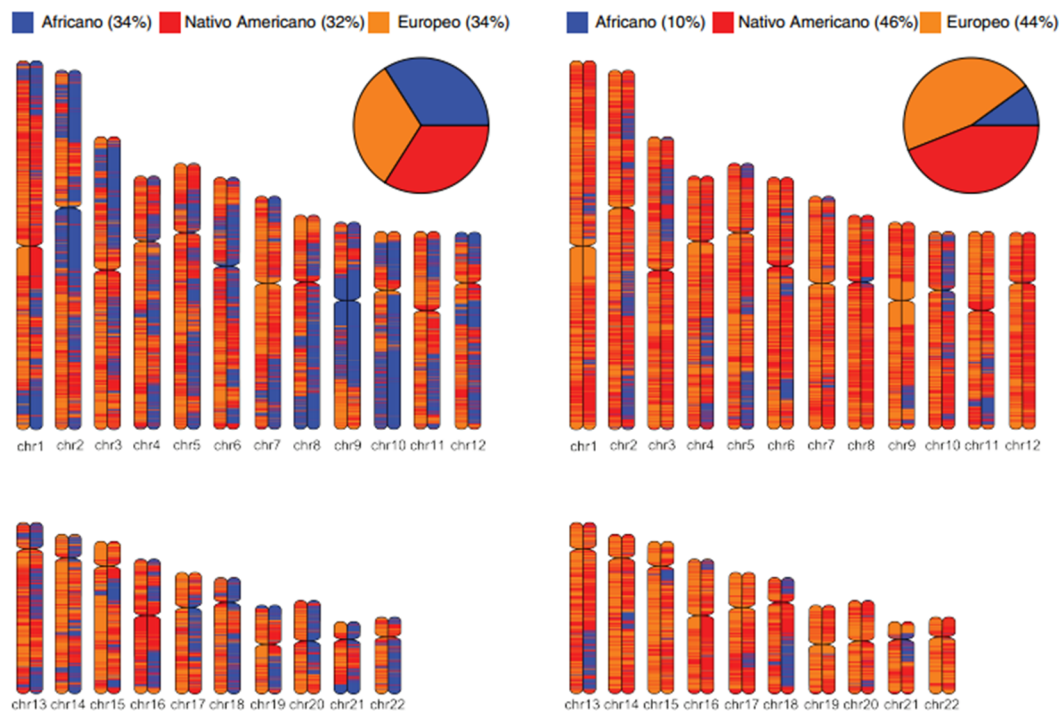


Figura 3.1: Regiones (locus) y el mestizaje genético ancestral en genomas de Colombia. Se muestran los cromosomas coloreados de dos individuos colombianos. Los colores muestran las distribuciones genómicas de loci con ancestralidad de África, Asia (nativos americanos) y Europa. Los pasteles describen las proporciones generales ancestrales en todo el genoma. Chr: cromosoma.

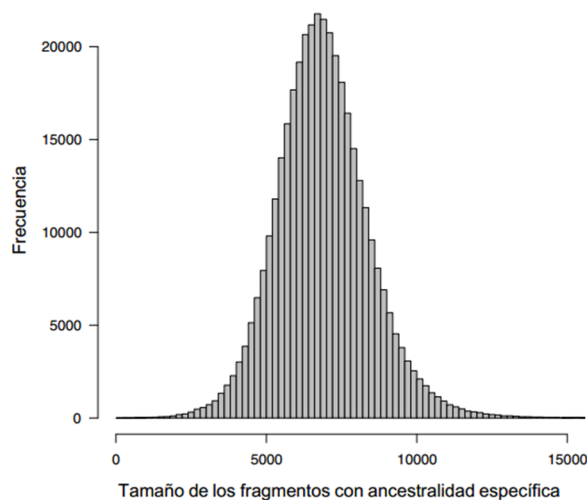


Figura 3.2: Distribución de los tamaños de los segmentos cromosómicos con ancestralidad específica usados para el análisis de enriquecimiento.

Como se puede observar en la figura 3.1, los individuos colombianos varían ampliamente con respecto a sus tres contribuciones ancestrales generales como en sus orígenes ancestrales regionales (de locus). Estos patrones de ancestralidad específica de locus de

los colombianos fueron explorados para buscar regiones genómicas irregulares que tuvieran contribuciones enriquecidas (más pronunciadas) de una población ancestral particular. Los límites de estos segmentos cromosómicos individuales (loci) fueron definidos por mapas de recombinación previamente descritos.¹⁶ Hay en total 379.218 de dichos loci en todo el genoma y el promedio de la longitud de los bloques de ancestralidad de los segmentos de ancestralidad fue de 7,542bp (ver figura 3.2).

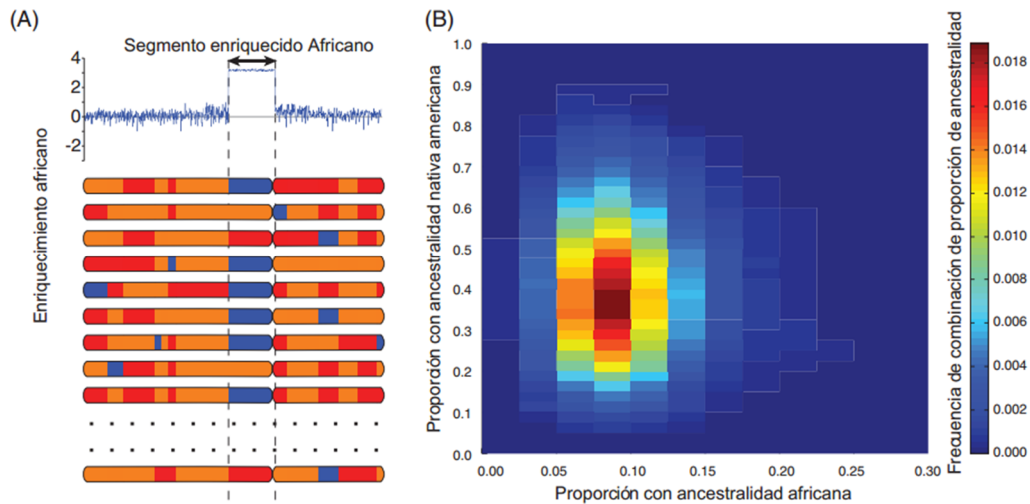


Figura 3.3: Análisis de enriquecimiento ancestral poblacional. A) Representación esquemática de la metodología. Las regiones ancestrales enriquecidas se identifican como loci genómicos con contribuciones ancestrales anormalmente altas para toda la población de alguno de los tres componentes continentales. B) Mapa de calor que muestra las frecuencias poblacionales de las combinaciones de ancestralidad triple para los genomas colombianos. Cada bloque corresponde a una combinación de proporciones ancestrales triples y la frecuencia de dicha combinación en la población se codifica con un color como se muestra en la escala del lado derecho. Es de notar que dado que se trata de combinaciones tres componentes ancestrales, el valor del tercer componente (europeo en este caso) es dependiente de los otros dos y por eso no se muestra.

La justificación para la aproximación de enriquecimiento ancestral se ilustra en la figura 3.3; los segmentos ancestrales enriquecidos son identificados en virtud de tener irregularmente altos niveles de contribución de un ancestro específico, comparado con los valores esperados según las proporciones de ancestralidad promedio de la población. Las frecuencias de las diferentes proporciones de combinaciones ancestrales en la población colombiana se presentan en la figura 3.4 como un mapa de calor. Las combinaciones de proporciones ancestrales más comunes se muestran en las regiones más calientes (rojo), mientras que las regiones frías (azul) muestran valores de combinaciones de proporciones ancestrales menos probables. Los segmentos cromosómicos en los que se localizan tales combinaciones de proporciones ancestrales anormales son los que resultan identificados con la prueba estadística de enriquecimiento ancestral aplicada aquí. Cuando la técnica de análisis de enriquecimiento ancestral es aplicada a los datos de las combinaciones de proporciones ancestrales de toda la población analizada, muchos segmentos cromosómicos enriquecidos (estadísticamente significativos) son revelados a través de todo el genoma (ver Figura 3.4).

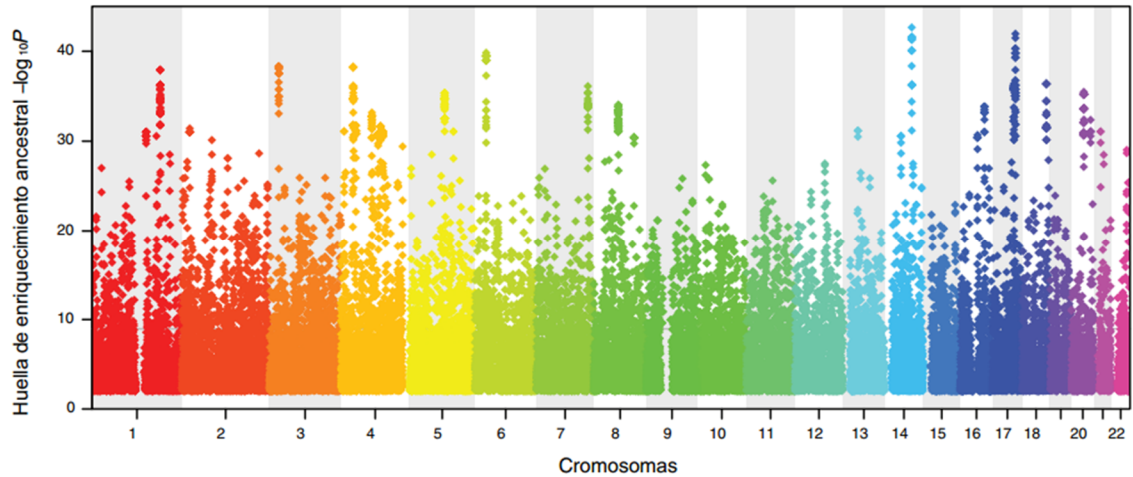


Figura 3.4: Gráfico de Manhattan mostrando las regiones genómicas identificadas como enriquecidas ancestralmente según la probabilidad trinomial (eje y) de observar una combinación ancestral particular en la población. El gráfico muestra el análisis para los autosomas.

Tabla 3.1: Genes localizados en segmentos cromosómicos enriquecidos, identificados en genomas colombianos

Cr	Inicio	Fin	Gen	Ancestralidad	-log10Prob
chr14	75179849	75203390	FCF1	Africana	41,32
chr14	75127954	75179807	AREL1	Africana	40,08
chr3	25831562	25836025	OXSM	Africana	38,31
chr3	25760434	25824989	NGLY1	Africana	38,09
chr17	58254690	58469586	USP32	Asiática	36,39

Cr: Cromosoma en el cual se localiza la región enriquecida ancestralmente. Inicio: Número de nucleótido sobre el cromosoma respectivo en el que inicia el segmento enriquecido. Fin: Número de nucleótido donde finaliza la región enriquecida. Gen: Nombre oficial, según HGNC, del (de los) gen(es) que se localiza(n) en las regiones enriquecidas. Ancestralidad: Identidad ancestral de la región sobrerrepresentada. -log10Prob: valor de significancia estadística con transformación log10.

En la tabla 3.1 se muestran los primeros cinco genes, de una lista total de más de 3000 genes localizados en los segmentos con mayor probabilidad de poseer un enriquecimiento anormal de ancestralidad.

Genes relacionados con salud y selección en las regiones de ancestralidad enriquecida: Una vez identificados los loci genómicos con niveles de contribución de ancestralidad irregulares, se procedió a investigar los genes localizados en estas regiones para saber si habían sido previamente implicados como evidencia de selección natural y si habían sido asociados con el proceso de salud enfermedad en humanos. Sólo los segmentos cromosómicos de ancestralidad enriquecida que tuvieron valores estadísticamente significativos ($P < 10^{-9}$, FDR valor q 6.5×10^{-9}) se utilizaron para estos análisis. Hay muchos genes en estas regiones que han sido identificados previamente como evidencia de selección positiva en poblaciones de alguna de las tres regiones ancestrales que contribuyeron a la población mestizada moderna de Colombia; hay también varios genes en regiones ancestralmente enriquecidas que han demostrado tener un rol importante en

varios fenotipos relacionados con enfermedad en estudios previos de selección y asociación (ver Tabla 3.2). En particular, hay varios genes relacionados al sistema inmune que han sido previamente caracterizados como sujeto de selección positiva en regiones ancestralmente enriquecidas (*CD226*, *HLA-B*, *MICA* and *MAPK10*).

Tabla 3.2: Genes localizados en regiones enriquecidas ancestralmente y sus características asociadas.						
Gen	Función*	Ancestralidad	Característica asociada	Evidencia	PMID	-log ₁₀ P
ADCY3	<i>Adenylate cyclase 3</i>	Asiática	Índice de Masa Corporal	PS	22344219	18.6
ATM	<i>Ataxia telangiectasia mutated</i>	Asiática	Ciclo celular.	AS	24390342 21983787 23419831	15.7
BCHE	<i>Butyrylcholinesterase</i>	Africana	Riesgo de enfermedad cardiovascular en europeos	AS	21943158 9780523	23.7
CASP8	<i>Caspase 8, apoptosis-related cysteine peptidase</i>	Asiática	Riesgo reducido en cáncer de mama	PS	17293864	9.6
CD226	<i>CD226 molecule</i>	Asiática	Adhesión celular, Sistema inmune adaptativo	AS	24390342 23128233 21829393 17554260	34.5
HLA-B	<i>Major histocompatibility complex, (MHC) class I, B</i>	Africana	Respuesta inmune	PS	23731540 16998491	31.6
MANBA	<i>Mannosidase</i>	Asiática	Riesgo reducido de Cáncer Colorectal y Asociado con Manosidosis	AS	17899454 21833088	29.7
MAPK10	<i>Mitogen-activated protein kinase 10</i>	Asiática	Sistema inmune	AS	25634076	25.4
MICA	<i>MHC class I polypeptide-related sequence A</i>	Africana	Respuesta inmune	PS	23731540	32.6
NANOS3	<i>Nanos homolog 3</i>	Asiática	Desarrollo de célula germinal	AS	21421998	23.9
NFKB1	<i>Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1</i>	Asiática	Riesgo reducido en cáncer de mama	PS	22562547	30.8
SF3B4	<i>Splicing Factor 3b, Subunit 4</i>	Asiática	Altura	AS	20881960 18391951	30.6
SLC24A5	<i>Solute carrier family 24, member 5</i>	Europea	Disminución de la pigmentación con Melanina en piel	PS	17182896	16.1
SLC44A2	<i>Solute carrier family 44, member 2</i>	Europea	Metabolismo	AS	22040064	20.1
ULBP1	<i>UL16 binding protein 1</i>	Europea	Vía de detección de ADN citosólico	AS	20923822	12.3
USP32	<i>Ubiquitin Specific Protease 32</i>	Asiática	Sobreexpresado en cáncer de mama	PS	19307593	36.4

*Las funciones aquí descritas se dejaron en inglés con el propósito de facilitar su búsqueda y localización. PS: *Positive selection*, selección positiva. AS: *Association study*, Estudio de asociación. -log₁₀Prob: valor de significancia estadística con transformación log₁₀.

El gen *HLA-B* localizado en el loci del Complejo Mayor de Histocompatibilidad (MHC) en el cromosoma 6 se encuentra en un segmento cromosómico enriquecido con ancestralidad de origen africano (ver Figura 3.5).

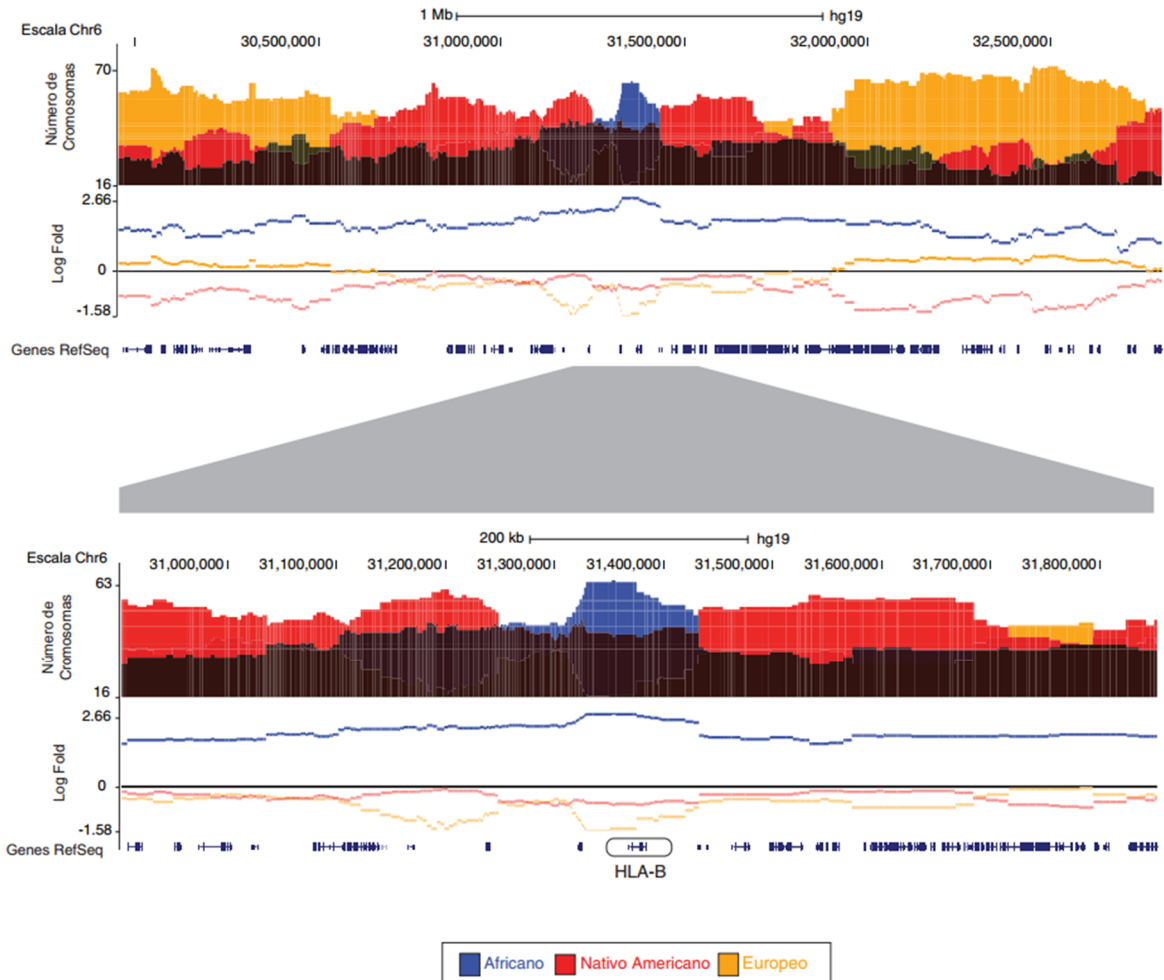


Figura 3.5: Ejemplo de una región con enriquecimiento ancestral africano hallada en el locus del HLA. Los recuentos cromosómicos poblacionales de los tres componentes ancestrales se muestran arriba del eje genómico y los valores de enriquecimiento (log fold – recuentos observados/recuentos promedio genómicos) para los componentes ancestrales se muestran abajo del eje. El panel superior muestra una región de ~2.5Mb del locus del HLA en el cromosoma 6 y el panel de abajo muestra una región ampliada del gen *HLA-B* enriquecido con ancestralidad africana (azul).

El *HLA-B* es un gen del MHC Clase I, que codifica una proteína de superficie celular que presenta los péptidos foráneos (antígenos) a las células del sistema inmune. El *HLA-B* es un gen altamente diverso con muchas variantes alélicas; la alta diversidad de los genes de *HLA* se le otorga a la habilidad de facilitar responder a un repertorio diverso de antígenos foráneos. En concordancia con esto, los niveles de diversidad en los genes de *HLA* entre las poblaciones humanas globales están positivamente correlacionados con los niveles específicos regionales de riqueza de patógenos, siendo las poblaciones africanas las que poseen el mayor de diversidad de alelos de *HLA*.¹² Adicionalmente, los niveles de

diversidad alélica del *HLA-B* alrededor del mundo muestran una fuerte correlación positiva con una presión selectiva de la malaria.¹⁷ Un enriquecimiento de alelos africanos en el locus *HLA-B* puede haber ayudado a la población colombiana a responder más efectivamente a la malaria y/u otros agentes patógenos. El gen *MAPK10* también está relacionado con el sistema inmune y está localizado en un segmento de ancestralidad principalmente asiática en la población colombiana. Recientemente, los alelos de este gen han sido identificados como sujeto de selección positiva para resistencia a la malaria en Malasia.¹⁸

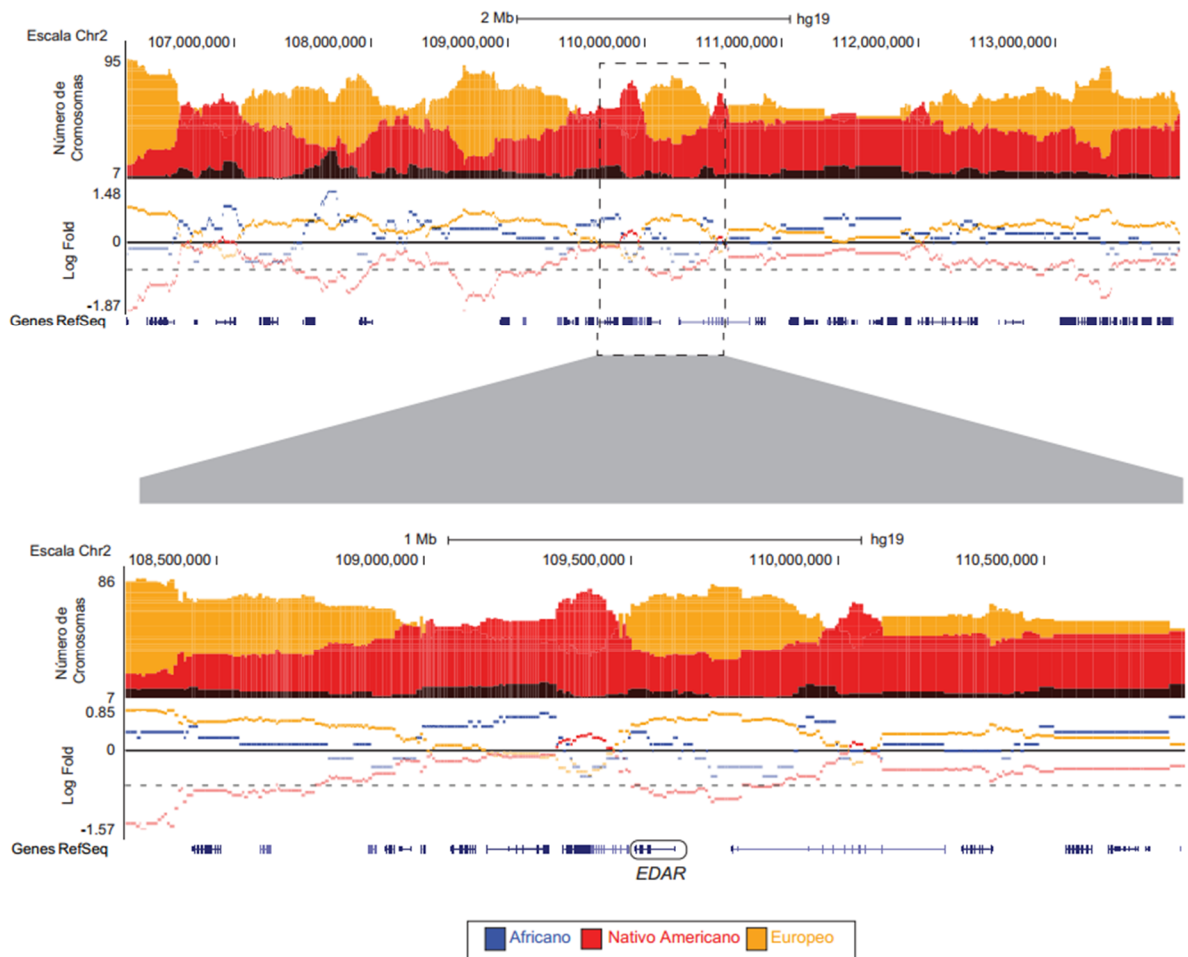


Figura 3.6: Ejemplo de una región con ancestralidad asiática (nativa americana) nula en la que se encuentra el locus del gen *EDAR*. Los recuentos cromosómicos poblacionales de los tres componentes ancestrales se muestran arriba del eje genómico y los valores de enriquecimiento (log fold – recuentos observados/recuentos promedio genómicos) para los componentes ancestrales se muestran abajo del eje. El panel superior muestra una región de ~7Mb rodeado el locus sobre el cromosoma 2 y el panel inferior muestra una visión ampliada del gen *EDAR* inmerso en una región sin ancestralidad asiática (nativa americana).

Otros genes interesantes localizados en regiones enriquecidas ancestralmente incluyen *SCL4A5* y *EDAR*. El en *SCL4A5* está localizado en una región genómica con ancestralidad enriquecida de origen europeo en la población colombiana. La selección positiva en este

locus ha sido previamente asociada con la evolución de menor pigmentación de melanina en la piel a medida que los europeos y los asiáticos se alejaron radialmente de África.⁷ El color de piel más claro ha sido propuesto como un efecto evolutivo por vía de selección sexual basado en las preferencias masculinas por parejas con colores de piel más claros,¹⁹ tal como originalmente fue propuesto por Darwin, aunque ésta hipótesis ha sido controvertida.²⁰ Es interesante especular si el enriquecimiento de ancestralidad europea en este locus puede ser atribuido a una preferencia por mujeres de piel clara en Colombia.

El gen *EDAR* por su parte ha recibido considerable atención recientemente debido a una conexión interesante el fenotipo que el codifica y evidencias de selección positiva específica de población.^{4,21} *EDAR* codifica un receptor de superficie celular que está involucrado en el desarrollo de folículos pilosos y glándulas cutáneas; y existe evidencia de selección positiva de alelos de *EDAR* en poblaciones asiáticas y nativas americanas. El alelo seleccionado reduce la evaporación de estructuras faciales expuestas y vías aéreas altas, lo que se piensa es una adaptación a los ambientes fríos y secos de Asia oriental. En la población de genomas mestizados colombianos, *EDAR* se encuentra en un segmento cromosómico donde la ancestralidad asiática (nativa americana) es nula (Figura 3.6). Esto podría reflejar el hecho de que la adaptación de *EDAR* a ambientes fríos y secos podría conferir una desventaja en los ambientes tropicales más cálidos y húmedos de Colombia.

Genes relacionados con la respuesta inmune, localizados en regiones con ancestralidad enriquecida: Adicionalmente al análisis dirigido descrito anteriormente para explorar genes localizados en regiones enriquecidas ancestralmente y que arrojaron resultados como los ejemplos ya descritos, también se usó un análisis de enriquecimiento funcional para evaluar algunas vías o cascadas específicas estaban sobrerrepresentadas entre los genes enriquecidos ancestralmente. Para ello se realizaron análisis de cascadas de grupos de genes enriquecidos (*Gene Set Enrichment Analysis*),¹⁵ de forma separada para loci genómicos que mostraran contribuciones ancestrales enriquecidas africanas, asiáticas (nativas americanas) y europeas. Un número de vías o cascadas resultaron sobrerrepresentadas cuando se comparaban contra genes enriquecidos ancestralmente (ver Figura 3.7A) y algunas vías similares se encuentran entre los diferentes componentes ancestrales a pesar del hecho de que sus genes componentes son excluyentes mutuamente por definición operacional. En particular, las cascadas relacionadas con la respuesta inmune estuvieron significativamente sobrerrepresentadas para todos los componentes ancestrales.

La sobrerrepresentación significativa de genes enriquecidos ancestralmente entre los componentes del sistema inmune, incluye genes que se localizan en cascadas de reacciones tanto de la respuesta inmune innata como de la adaptativa (ver Figura 3.7B).

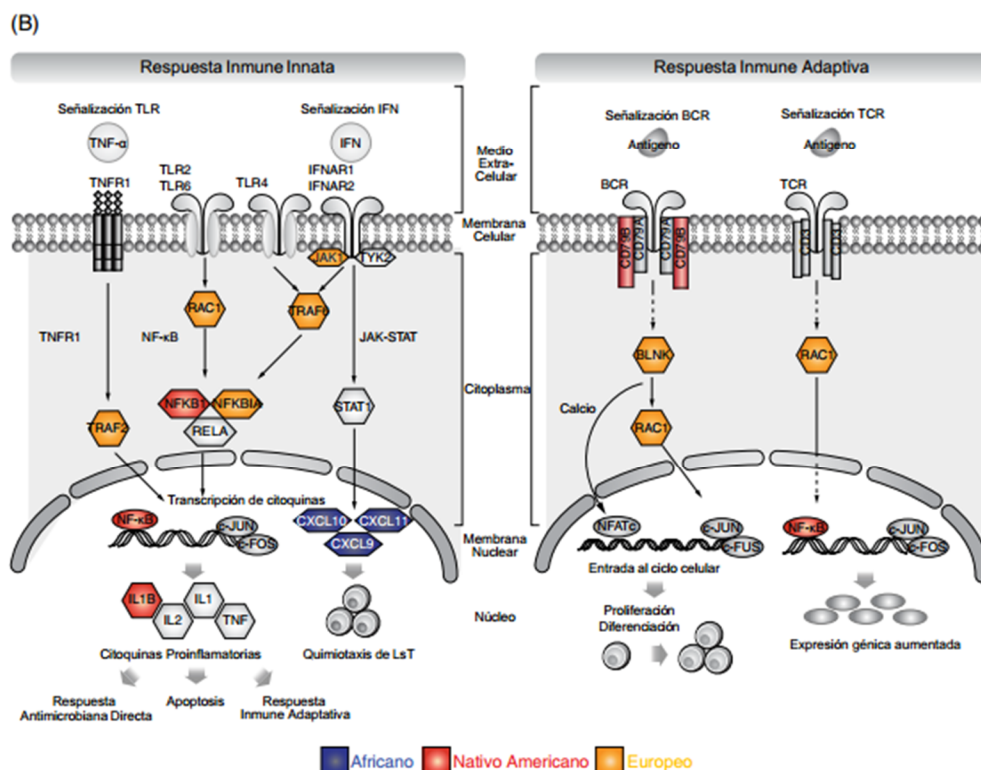
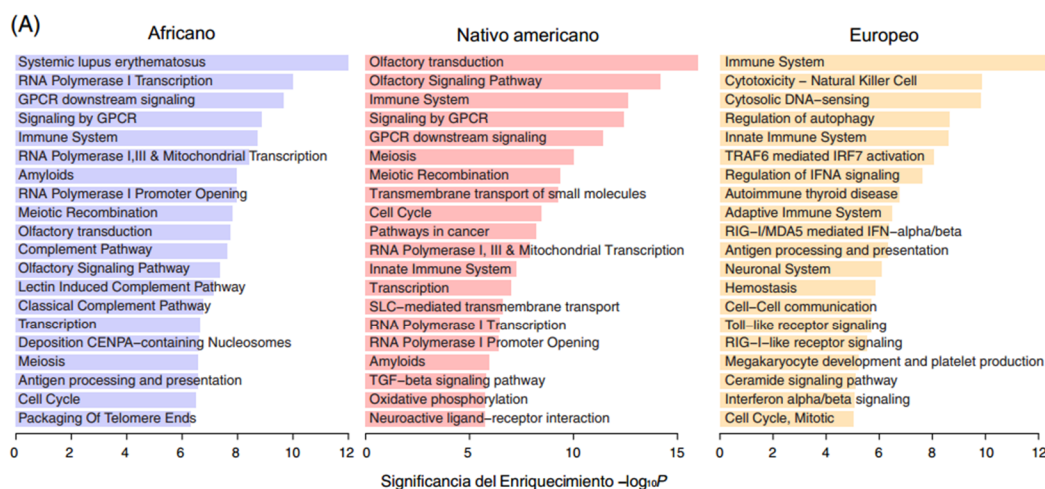


Figura 3.7: Análisis de enriquecimiento funcional en regiones con ancestralidad enriquecida. A) Vías (cascadas) identificadas como sobrerrepresentadas entre las regiones con ancestralidad enriquecida según análisis GSEA. B) Esquema de 4 de esas vías que están involucradas en la respuesta inmune innata y adaptativa.

Las vías, relacionadas entre sí, de los TLR¹³ y del Interferón de la respuesta inmune innata contienen numerosos genes con enriquecimiento ancestral, incluyendo miembros citoplasmáticos de las cascadas de señalización de NF- κ B y JAK-STAT. Existen también genes enriquecidos ancestralmente que codifican miembros de estas cascadas, incluyendo factores de transcripción (NF- κ B), citoquinas inflamatorias (IL1B) y quimiocinas (CXCL9, 10

¹³ Toll-like Receptor

and 11), las cuales en conjunto ayudan a direccionar la respuesta antimicrobiana a través de mecanismos tales como la apoptosis y la quimiotaxis de células T. Los genes enriquecidos ancestralmente que están involucrados en la respuesta inmune adaptativa codifican miembros de las vías de señalización de los receptores células de células T y B, incluyendo receptores de antígenos (CD79B) y moléculas de señalización citoplasmática (BLNK y RAC1) junto con el factor de transcripción NF- κ B. Estas proteínas ayudan a facilitar la proliferación y la diferenciación de células T y B en respuesta a retos específicos al sistema inmune. Llamativamente, estas vías contienen una mezcla de genes con enriquecimiento ancestral a partir de diferentes componentes ancestrales. Esto indica que los individuos mestizados genómicamente de Colombia han ensamblado vías del sistema inmune que están hechas de combinaciones de alelos con ancestralidad específica que nunca se han visto en contacto con el mismo trasfondo genético.

Esto puede haber provisto un mecanismo para confrontar las nuevas combinaciones de patógenos microbianos encontrados en el nuevo mundo. Se debe anotar que la función génica tiende a ser espacialmente correlacionada y esto es particularmente cierto para el tipo de genes del sistema inmune discutidos aquí. Dado que estos conceptos pueden ser interpretados como una violación a los supuestos de independencia que subyacen al análisis de vías tipo GSEA, los valores *P* reportados aquí deben ser interpretados con cuidado.

Discusión

Los resultados del tipo de análisis propuesto aquí indican que existe un enriquecimiento sustancial específico de ancestralidad en diversos loci a través de todo el genoma y que estas regiones codifican numerosos genes involucrados en funciones relacionadas al sistema inmune así como genes previamente implicados en selección natural para pigmentación de la piel y desarrollo glandular.

Mestizaje genómico, adaptación al medio y selección en Colombia: Cuando se comparan los patrones de contribuciones ancestrales específicos a través de toda la población de genomas colombianos estudiados aquí, es muy claro que los patrones de mestizaje no ocurren al azar (ver Figura 3.3). Hay numerosos segmentos cromosómicos enriquecidos que tienen niveles anormalmente altos o bajos de ancestralidad de algunas de las tres poblaciones ancestrales. Estos resultados sugieren la posibilidad de que los loci enriquecidos hayan sido diferencialmente retenidos en la población moderna de Colombia debido a su utilidad en el medio ambiente.

Aquí se propone que el proceso particular de "selección" que se produjo en la población mestizada de Colombia, basado en la variedad de alelos ancestrales específicos preexistentes, fue de alguna manera distinto del concepto de selección natural como normalmente se propone. La selección natural adaptativa, también conocida como positiva o direccional, es frecuentemente considerada como el resultado de la fijación de

nuevas mutaciones basadas en el éxito reproductivo diferencial. Dicho proceso inicia con la introducción de nuevos alelos por mutación a una frecuencia poblacional muy baja y por lo tanto es relativamente lento, de tal forma que típicamente toma decenas o cientos de miles de años.²² El proceso evolutivo que da lugar a los segmentos enriquecidos ancestralmente en la población mestizada de Colombia fue por el contrario basado en selección entre alelos preexistentes de poblaciones específicas. Estos alelos específicos de población (o enriquecidos) evolucionaron *in situ* en sus regiones ancestrales, basados en presiones selectivas locales, sobre el curso de muchos miles de años. Cuando las poblaciones que contenían estos alelos llegaron al nuevo mundo, los alelos adaptados preexistentes estaban listos para ser rápidamente redistribuidos en nuevos genomas mezclados (mestizados). En otras palabras, el largo y lento proceso de selección natural había ya ocurrido en las poblaciones fuentes ancestrales para generar un grupo actualizado de variaciones genéticas con una amplia variedad de utilidades adaptativas. A partir de este grupo ya existente de variación genética, numerosos segmentos con ancestralidad específica que contenían alelos adaptados estaban asequibles para ser redistribuidos. En este sentido, el enriquecimiento de segmentos con ancestralidad específica en la población colombiana, basado en su utilidad en el medio ambiente del nuevo mundo, podría haber sucedido mucho más rápidamente, es decir dentro del espacio de tiempo relativamente corto que siguió después de la llegada de Cristóbal Colón a América. La capacidad de rápidamente formar genomas mezclados (mestizados) sobre un tiempo relativamente corto, con combinaciones irregulares de alelos específicos ancestrales, se apoya en patrones asimétricos del mestizaje específico de sexo observado en las poblaciones latinoamericanas.

Una estrategia similar de búsqueda de evidencia de selección basada en enriquecimiento ancestral ha sido desarrollada en dos estudios recientes de poblaciones americanas y africanas mezcladas.^{23,24} Los resultados de estos estudios difieren drásticamente. El primer estudio reportó seis regiones con niveles altamente significativos de enriquecimiento ancestral.²³ El estudio más reciente incluyó un tamaño de muestra sustancialmente mayor así como un control astringente para múltiples pruebas estadísticas y no fue capaz de replicar los hallazgos previos de regiones genómicas significativamente enriquecidas ancestralmente.²⁴ Una de las limitaciones potenciales de la estrategia utilizada en este trabajo es el limitado tamaño de muestra comparado con estudios previos que han empleado genotipos obtenidos por microarreglos (SNP chips). Es posible que los resultados de las regiones identificadas como enriquecidas ancestralmente en este trabajo no logren ser replicados cuando muestras de mayor tamaño sean utilizadas, como en el caso de los dos estudios previamente mencionados con afroamericanos. Sin embargo, nuestra estrategia tiene la ventaja de una resolución incrementada para la inferencia de la ancestralidad local que provee la comparación de genomas completos, la cual también necesita de tamaños relativamente bajos comparado a los estudios de genotipos. En este trabajo además se utilizó una estrategia estadística diferente para la identificación de las regiones de ancestralidad enriquecida a la utilizada

en estudios previos, la cual permitió buscar desviaciones de los patrones ancestrales triples esperados. Esta estrategia se aplicó utilizando controles astringentes para múltiples pruebas estadísticas. Es de valor también hacer notar que la población colombiana analizada aquí, y las poblaciones latinoamericanas en general, muestran más mestizaje genómico que lo observado para poblaciones afroamericanas en los estudios previamente mencionados. Es posible que niveles más elevados de mestizaje provean una mejor oportunidad para la retención diferencial de alelos ancestrales previamente seleccionados.

Mestizaje genómico y la carga de morbilidad infecciosa en Colombia: Los roles funcionales de los genes localizados en las regiones genómicas mezcladas anormalmente sugiere que la retención de alelos con ancestralidad específica en la población colombiana está más estrechamente ligada a la función del sistema inmunológico (ver Tabla 3.2 y Figura 3.7). Este proceso es probable que haya sido impulsado por la exposición de la población mezclada a numerosos agentes patógenos microbianos, muchos de los cuales nunca antes habían sido enfrentados. Estos nuevos agentes patógenos podrían haberse originado en el ambiente tropical hallado en el nuevo mundo y/o podrían ser el resultado de la mezcla de los tres grupos de poblaciones continentales diferentes, cada una con su propio conjunto de patógenos endémicos.

Colombia se encuentra en la región tropical, es atravesada por la línea ecuatorial y tiene una alta carga de enfermedades infecciosas, causadas por una variedad de agentes patógenos. En comparación con otros países de la región y en todo el mundo, Colombia tiene un alto nivel de riqueza de patógenos, tal como se evidencia por el número de agentes patógenos conocidos en el país ($n=244$).^{17,45} La carga de las enfermedades infecciosas en Colombia incluye las transmitidas por agua, como el cólera, y las transmitidas por vectores, como la fiebre amarilla, el dengue y la malaria.²⁵ Los factores genéticos del hospedero juegan un papel importante en la disposición de la susceptibilidad y la resistencia de los individuos a estos agentes patógenos.¹³ Se ha demostrado previamente que la selección natural aumenta la frecuencia de los alelos que proporcionan una mayor resistencia al cólera, el dengue y la malaria. Además, muchos de estos alelos son específicos de poblaciones particulares (o enriquecidos), habiendo evolucionado dentro de tales poblaciones en las regiones geográficas en las que se originaron las enfermedades.

La susceptibilidad al cólera, que es endémica en el sudeste de Asia, se sabe que está fuertemente influenciada por factores genéticos del huésped.^{26,27} Estos incluyen genes que codifican miembros de la vía NF- κ B que funciona como parte del sistema inmune innata.^{49,50} Los resultados de nuestro análisis muestran como existen genes de esta vía en regiones con patrones anómalos de ancestralidad, incluyendo genes que se encuentran dentro de segmentos con enriquecimiento de origen asiático (nativos americanos) (ver Figura 3.7).

Existen numerosos estudios que han identificado factores genéticos del huésped que parecen atenuar la susceptibilidad al dengue.²⁸ Estos incluyen estudios de asociación de alcance genómico (GWAS¹⁴),²⁹⁻³¹ como estudios de genes candidatos a menor escala.³²⁻³⁴ Un estudio GWAS reciente encontró los SNPs asociados más significativamente con el dengue en el locus *MICB*, así como otros SNPs adicionales en los locus adyacentes de *HLA-B* y *HLA-C*.²⁹ Todos estos genes son miembros del locus MHC que codifica numerosas proteínas de los receptores de superficie celular involucrados en la respuesta inmune a través de la presentación de secuencias de péptidos antigénicos. Los tres genes se encuentran dentro de segmentos genómicos enriquecidos con ancestralidad africana (ver Tabla 3.2). Otros estudios de genes candidatos también han descubierto asociaciones entre alelos específicos de HLA y la susceptibilidad a dengue en poblaciones asiáticas.³²⁻³⁴

A la fecha no se conocen alelos de resistencia del hospedero asociados con la susceptibilidad a la fiebre amarilla. Sin embargo, durante mucho tiempo se ha pensado que las poblaciones con diferentes ancestralidades genéticas tienen diferentes grados de susceptibilidad a la fiebre amarilla, lo que sugiere un componente genético en la susceptibilidad a este patógeno. En particular, se pensaba que las poblaciones con ancestralidad africana poseen una susceptibilidad reducida a la fiebre amarilla,³⁵⁻³⁷ y esta noción se usó para estimular un mayor incentivo económico por el uso de mano de obra esclava africana en el nuevo mundo.³⁸ Sin embargo, esta idea ha sido criticada con base en la falta de consenso histórico y la ausencia de una evidencia directa de un componente ancestral genético en la resistencia, tal como si se ha observado en la malaria.³⁹ Sin embargo, un análisis reciente de la mortalidad por fiebre amarilla en el siglo 19 en los Estados Unidos proporcionó evidencia estadística convincente para una asociación entre la ancestralidad genética y la susceptibilidad a la fiebre amarilla.⁴⁰ Se encontró que las personas con ancestralidad africana tenían tasas de incidencia de fiebre amarilla muy similares en comparación con los de origen europeo, pero tenían tasas significativamente más bajas de mortalidad. Los autores trataron de descartar una serie de factores ambientales, sociales y/o económicos que lograran explicar esta diferencia y concluyeron que sus resultados apoyan la existencia de factores genéticos del huésped que aminoran la gravedad de las infecciones de fiebre amarilla entre las personas de ancestralidad africana.

La malaria es una amenaza particularmente grave y constante para la salud pública en Colombia, con patrones de transmisión tanto endémicos como epidémicos.⁴¹ Mientras que Medellín no tiene una alta incidencia de la malaria debido a su altitud, esta enfermedad sí es endémica de las regiones cercanas a altitudes más bajas, sobre todo en áreas de la selva a lo largo de las costas Atlántica y el Pacífica.⁴² Dado que la población de Medellín es cosmopolita,⁴³ habiendo recibido constantemente inmigrantes de las regiones

¹⁴ Genome Wide Association Study

circundantes a través de los siglos, la malaria puede haber ejercido una presión selectiva sobre la población. Numerosos alelos en diferentes locus genéticos han evolucionado bajo la influencia de una presión selectiva sobre la base de la resistencia a la malaria.¹³ Este proceso se ha producido de forma independiente en diferentes poblaciones de todo el mundo y ha dado lugar a una serie de alelos de resistencia con ancestralidad específica. Por ejemplo, existen distintos alelos específicos de África y Asia que codifican por la cadena beta de hemoglobina (HBB) que han evolucionado de forma independiente y proveen resistencia a la malaria.⁴⁴⁻⁴⁶

Existe una clara relación entre la carga de morbilidad infecciosa en Colombia, para algunos de los agentes patógenos más comunes que se encuentran en el país, como se ha descrito anteriormente, y factores genéticos del huésped que median en la susceptibilidad y resistencia. Muchos de estos factores genéticos del huésped es probable que sean específicos de poblaciones ancestrales (o enriquecidos) que se desarrollaron inicialmente dentro de las poblaciones de origen ancestral. Los resultados obtenidos en nuestro análisis enriquecimiento ancestral sugieren que la población colombiana moderna puede haber tenido la oportunidad de extraer, a partir de una variedad de genes con ancestralidad específica, un repertorio de alelos de resistencia para ensamblar un grupo de genes relacionados con el sistema inmune capaz de contrarrestar las amenazas de la amplia variedad de patógenos que se encuentran en el entorno del nuevo mundo. Además de ser de interés desde una perspectiva evolutiva, estos resultados pueden tener implicaciones para comprender mejor los determinantes genéticos de la salud en las poblaciones altamente mezcladas de América Latina.

Referencias capítulo 3

1. Pickrell, J.K., *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome research* 19, 826-837 (2009).
2. Sabeti, P.C., *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913-918 (2007).
3. Sabeti, P.C., *et al.* Positive natural selection in the human lineage. *Science* 312, 1614-1620 (2006).
4. Grossman, S.R., *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* 152, 703-713 (2013).
5. Williamson, S.H., *et al.* Localizing recent adaptive evolution in the human genome. *PLoS genetics* 3, e90 (2007).
6. Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human AIM1 locus in a European population. *Molecular biology and evolution* 23, 179-188 (2006).
7. Norton, H.L., *et al.* Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular biology and evolution* 24, 710-722 (2007).
8. Myles, S., Somel, M., Tang, K., Kelso, J. & Stoneking, M. Identifying genes underlying skin pigmentation differences among human populations. *Human genetics* 120, 613-621 (2007).
9. McEvoy, B., Beleza, S. & Shriver, M.D. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Human molecular genetics* 15 Spec No 2, R176-181 (2006).

10. Lao, O., de Gruijter, J.M., van Duijn, K., Navarro, A. & Kayser, M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human genetics* 71, 354-369 (2007).
11. Izaguirre, N., Garcia, I., Junquera, C., de la Rua, C. & Alonso, S. A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular biology and evolution* 23, 1697-1706 (2006).
12. Sanchez-Mazas, A., Lemaitre, J.F. & Currat, M. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367, 830-839 (2012).
13. Karlsson, E.K., Kwiatkowski, D.P. & Sabeti, P.C. Natural selection and infectious disease in human populations. *Nature reviews. Genetics* 15, 379-393 (2014).
14. Welter, D., *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42, D1001-1006 (2014).
15. Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102, 15545-15550 (2005).
16. Omberg, L., *et al.* Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC genetics* 13, 49 (2012).
17. Garamszegi, L.Z. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. *Malaria journal* 13, 349 (2014).
18. Liu, X., *et al.* Differential positive selection of malaria resistance genes in three indigenous populations of Peninsular Malaysia. *Human genetics* (2015).
19. Aoki, K. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Annals of human biology* 29, 589-608 (2002).
20. Madrigal, L. & Kelly, W. Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *American journal of physical anthropology* 132, 470-482 (2007).
21. Kamberov, Y.G., *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152, 691-702 (2013).
22. Garrigan, D. & Hammer, M.F. Reconstructing human origins in the genomic era. *Nature reviews. Genetics* 7, 669-680 (2006).
23. Jin, W., *et al.* Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome research* 22, 519-527 (2012).
24. Bhatia, G., *et al.* Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American journal of human genetics* 95, 437-444 (2014).
25. CIA. CIA World Factbook: Colombia. (2014).
26. Harris, J.B., *et al.* Susceptibility to *Vibrio cholerae* infection in a cohort of household contacts of patients with cholera in Bangladesh. *PLoS neglected tropical diseases* 2, e221 (2008).
27. Barua, D. & Paguio, A.S. ABO blood groups and cholera. *Annals of human biology* 4, 489-492 (1977).
28. Stephens, H.A. HLA and other gene associations with dengue disease severity. *Current topics in microbiology and immunology* 338, 99-114 (2010).
29. Khor, C.C., *et al.* Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nature genetics* 43, 1139-1141 (2011).
30. Whitehorn, J., *et al.* Genetic variants of MICB and PLCE1 and associations with non-severe dengue. *PloS one* 8, e59067 (2013).
31. Devignot, S., *et al.* Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PloS one* 5, e11671 (2010).
32. Stephens, H.A., *et al.* HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. *Tissue antigens* 60, 309-318 (2002).

33. Nguyen, T.P., *et al.* Protective and enhancing HLA alleles, HLA-DRB1*0901 and HLA-A*24, for severe forms of dengue virus infection, dengue hemorrhagic fever and dengue shock syndrome. *PLoS neglected tropical diseases* 2, e304 (2008).
34. Appanna, R., Ponnampalavanar, S., Lum Chai See, L. & Sekaran, S.D. Susceptible and protective HLA class 1 alleles against dengue fever and dengue hemorrhagic fever patients in a Malaysian population. *PloS one* 5(2010).
35. Kiple, K.F. *The Caribbean slave: a biological history*, (Cambridge University Press Cambridge, Cambridge, 1984).
36. Kiple, K.F. & Kiple, V.H. Black yellow fever immunities, innate and acquired, as revealed in the American South. *Social Science History* 1, 419-436 (1977).
37. McNeill, J.R. Ecology, epidemics and empires: environmental change and the geopolitics of tropical America, 1600-1825. *Environment and History* 5, 175-184 (1999).
38. McNeill, J.R. *Mosquito Empires: ecology and war in the Greater Caribbean, 1620-1914*, (Cambridge University Press, Cambridge, 2010).
39. Espinosa, M. The question of racial immunity to Yellow Fever in history and historiography. *Social Science History*.
40. Blake, L.E. & Garcia-Blanco, M.A. Human genetic variation and yellow fever mortality during 19th century U.S. epidemics. *mBio* 5, e01253-01214 (2014).
41. Ramirez, A.P., Buitrago, J.I., Gonzalez, J.P., Morales, A.H. & Carrasquilla, G. Frequency and tendency of malaria in Colombia, 1990 to 2011: a descriptive study. *Malaria journal* 13, 202 (2014).
42. Valero, M.V., Vergara, J. & Herrera, S. Malaria Atlas Project. (2015).
43. DANE. Censo General 2005: Perfil Medellin Antioquia. (2005).
44. Ohashi, J., *et al.* Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *American journal of human genetics* 74, 1198-1208 (2004).
45. Modiano, D., *et al.* Haemoglobin C protects against clinical Plasmodium falciparum malaria. *Nature* 414, 305-308 (2001).
46. Agarwal, A., *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* 96, 2358-2363 (2000).

Capítulo 4

El Chocó Colombia: un *hotspot*¹⁵ de biodiversidad humana

Resumen

Introducción: El Chocó es un departamento situado en la costa del pacífico colombiano que se caracteriza por tener una gran mayoría de población afrodescendiente. Dados los resultados obtenidos en los dos capítulos anteriores de esta tesis en una población con ancestralidad predominantemente europea, el objetivo de esta parte del estudio fue caracterizar la ancestralidad, el mestizaje genómico y la diversidad en una población con ascendencia claramente diferente y en este caso el grupo poblacional ideal era el del Chocó colombiano. **Metodología:** La variabilidad genética se caracterizó en una muestra de 101 donantes voluntarios (61 mujeres y 40 hombres) del departamento del Chocó. Los genotipos se determinaron de manera individual a través de la caracterización de 610,545 SNPs en todo el genoma. Los haplotipos para el genoma uniparental, mtADN (línea materna) y Y-DNA (línea paterna) también fueron determinados. Estos datos fueron utilizados para análisis comparativo con varias poblaciones de todo el mundo, incluidas las poblaciones ancestrales putativas de África, América y Europa, además de varias poblaciones mestizadas americanas. **Resultados:** La población del Chocó tiene ascendencia genética predominantemente africana (75,8%), con similares proporciones de contribución de origen europeo (13,4%) y de nativos americanos (11,1%). La población del Chocó muestra niveles relativamente altos una mezcla genética triple y niveles mucho más altos de ancestralidad nativa americana, en comparación con otras poblaciones del afrodescendientes del nuevo mundo (Caribe y Estados Unidos). Existe un patrón muy llamativo de ancestralidad específica de sexo en el Chocó, con aportes predominantes de ancestralidad nativa americana a lo largo de la linaje materno y de origen europeo a lo largo de la linaje paterno. La población del Chocó también se caracterizó por niveles relativamente altos de diversidad genética global en comparación con las poblaciones ancestrales putativas y otras poblaciones mestizadas americanas. **Conclusión:** Estos resultados sugieren un patrimonio genético único para la población del Chocó y ponen de relieve la profunda diversidad genética humana que se pueden encontrar en la región.

Introducción

El Chocó es un departamento situado a lo largo de la costa pacífica colombiana (ver Figura 4.1A). El Chocó se extiende de la frontera con Panamá en el norte hasta los límites con el valle del Cauca, en el suroeste de Colombia. El estado de Chocó es reconocido mundialmente como *hotspot*¹⁶ de la biodiversidad.

Un *hotspot* de biodiversidad se define como una región geográfica específica con una gran cantidad de biodiversidad endémica que se ve amenazada por la actividad humana.¹ Para que un área sea reconocida oficialmente como un *hotspot* de biodiversidad, al menos el 0,5% de sus especies de plantas vasculares (o 1.500 especies) debe ser caracterizado como endémicas, lo cual se define como especies que se encuentran únicamente dentro de una

¹⁵ Para esta palabra no existe una traducción literal ni cercana, se refiere a la concentración inusualmente alta de alguna característica. Para propósitos de esta tesis se adoptará el término *hotspot*.

¹⁶ <http://www.eoearth.org/view/article/150631/>

región geográfica o un tipo de hábitat. Existen 25 regiones en el mundo que cualifican como *hotspots* de biodiversidad de acuerdo con este criterio y en conjunto estas áreas son el hogar de al menos el 60% de las especies conocidas en el mundo de plantas, aves, reptiles, anfibios y mamíferos. Toda el área del Chocó está contenida dentro del llamado *hotspot* de biodiversidad del Chocó Darién, también conocido como el *hotspot* de biodiversidad Tumbes-Chocó-Magdalena o Chocó Biogeográfico¹⁷.



Figura 4.1: Biodiversidad humana en el Chocó. A) Departamento del Chocó sobre la costa pacífica colombiana. B) La gente del Chocó.

El *hotspot* de biodiversidad del Chocó Darién se extiende a lo largo de la costa pacífica, desde la región del canal de Panamá en el norte, a través de la selva del Darién y los bosques húmedos tropicales del Chocó en Panamá y Colombia, atravesando completamente el Ecuador antes de terminar en los bosques secos de la costa norte de Perú. Esta zona abarca una amplia variedad de hábitats, incluyendo los bosques lluviosos tropicales más húmedos del mundo que se encuentran en El Chocó, Colombia. La parte de Colombia de este *hotspot* de biodiversidad en El Chocó está relativamente preservada, en comparación con la zona ecuatoriana donde el 98% del bosque nativo ha sido deforestado. El *hotspot* de biodiversidad del Chocó Darién sustenta ~10.000 especies de plantas vasculares, junto con 600 especies de aves, 235 especies de mamíferos, 350 especies de anfibios y 210 especies de reptiles.

Aunque El Chocó es ampliamente reconocido como un *hotspot* de la biodiversidad, un aspecto crítico de la biodiversidad de la región sigue siendo en gran parte inexplorado - la diversidad de su población humana (ver Figura 4.1B). La población de El Chocó tiene una herencia genética única, principalmente africana con mezcla de las Américas y Europa. La gran mayoría de la población es afrocolombiana (82,1%), pero también hay un número considerable de los nativos americanos (12,7%) y de personas con ancestralidad

¹⁷ <http://tmalliance.org/about/where-we-work/biodiversity-hotspot/>

principalmente europea (5,2%).² En Colombia existen entre 9 a 20 millones de afrodescendientes, por lo que es el país con la tercera mayor cantidad de afrodescendientes en América. A pesar de la presencia de una gran población de afrodescendientes en Colombia, existen muy pocos estudios genéticos de los afrocolombianos.³ Estudios previos de ancestralidad genética en colombianos, se han ocupado principalmente de genomas mestizados que tienen componentes primeramente europeo y en segundo lugar nativo americana.⁴⁻¹¹ Dado el alto porcentaje de afrocolombianos que viven en el Chocó, los estudios genéticos de esta población son ideales para descubrir la dimensión africana aún sin explotar de ascendencia colombiana y la biodiversidad humana.

Los colombianos, como se ha mostrado en los dos capítulos anteriores, tienen patrones de mestizaje genómico de tres orígenes, que resultan de las contribuciones ancestrales de África, América y Europa a la población moderna.⁹ La historia de ancestralidad de cada colombiano, junto con sus patrones de mestizaje genómico específico, se encuentra escrita en la secuencia de su genoma. Por lo tanto, el análisis de la secuencia del genoma se puede usar para inferir la ancestralidad genética y los patrones de mestizaje genómico para los individuos y para la población como un todo. A nivel individual, las proporciones generales de ancestralidad y mestizaje genómico de origen africano, nativo americano y europeo se pueden inferir tanto con patrones específicos de locus y de sexo. De la misma manera, la ancestralidad genética se puede explorar tanto a nivel continental para descubrir las regiones más amplias de origen de los colombianos y a un nivel subcontinental para explorar las regiones y países ancestrales específicos a partir de las cuales proceden los ancestros de cada individuo. A nivel de la población general, los patrones de mezcla de locus específicos pueden revelar si la selección natural ha enriquecido alguno de los ancestros específicos a lo largo de segmentos genómicos particulares. La inferencia a nivel poblacional también se puede ser utilizada para evaluar si existen diferencias específicas de sexo en la ancestralidad que resultan de contribuciones ancestrales no equilibradas en los linajes paternos y maternos. Este tipo de enfoques de investigación genómica se aplicará, con el fin de descifrar los patrones de ancestralidad genética y la mestizaje genómico de la población del Chocó.

Metodología

Panorama general: Muestras de saliva fueron utilizadas para aislar ADN genómico a partir de 101 voluntarios del departamento del Chocó. Las muestras de ADN fueron utilizadas para caracterizar la secuencia genética en 610.545 SNPs¹⁸. La colección completa de SNPs de un individuo es referida aquí como un genotipo. La identidad específica de los residuos de ADN de una secuencia dada en un grupo de SNPs ligados genéticamente se denomina haplotipo. Los genotipos fueron establecidos para todos los autosomas y los haplotipos para los ADN uniparentales: mtADN para linaje materno y Y-ADN para linaje paterno. Los

¹⁸ SNP: Single Nucleotide Polymorphism

genotipos del Chocó fueron comparados con genotipos de una variedad de poblaciones humanas (ver Tabla 4.1 en sección resultados) con objeto de hacer inferencias acerca de su ancestralidad y diversidad genética. Los haplotipos mtADN y Y-ADN fueron comparados con distribuciones globales conocidas de estos cromosomas con objeto de hacer inferencias acerca de los respectivos linajes, materno y paterno.

Donantes de muestras y determinación del genotipo: Donantes voluntarios de muestras de ADN fueron reclutados en la Universidad Tecnológica del Chocó, localizada en la ciudad de Quibdó. En total se reclutaron 101 voluntarios (61 mujeres y 40 hombres). Para el aislamiento de ADN se recolectaron muestras de saliva (método no invasivo) usando el kit Oragene®-Discover. Los donantes completaron un cuestionario acerca de autoidentificación étnica y origen familiar. Las muestras de ADN de los donantes fueron genotipificadas utilizando el chip de SNPs HumanOmniExpress-24 (Illumina).

Fuentes de los datos para genómica comparativa: Los genotipos de los donantes fueron comparados con datos de secuencias genómicas completas del Proyecto 1000 Genomas (1000G) y con datos de genotipificación del Proyecto de Diversidad del Genoma Humano (HGDP)¹⁹.¹²⁻¹⁵ Los genotipos de los donantes, así como como los genotipos de 1000G y PDGH, fueron mapeados en el espacio de coordenadas de la versión GRCh37/hg19 de referencia de la secuencia del genoma humano de Febrero de 2009 para los subsecuentes análisis.^{16,17} El programa PLINK fue utilizado para el control de calidad de genotipo y para extraer las posiciones autosómicos genotipificadas (es decir, los SNPs) comunes a las tres fuentes de genotipo para producir un conjunto de datos de genotipo fusionada final.¹⁸ Para el control de calidad, solo aquellas posiciones de SNPs con una tasa de genotipificación de $\geq 98\%$ se conservaron para los análisis.

Inferencia de ancestralidad y mestizaje genómico: El programa PLINK fue utilizado para escindir el conjunto de datos de genotipo combinado final mediante la eliminación de conjuntos de SNPs correlacionados. Las distancias genómicas fueron calculadas como distancias compartidas entre alelos apareados entre todos los genotipos escindidos individuales usando PLINK. La matriz resultante de distancias apareadas se proyectó en un gráfico bidimensional con el análisis de componentes principales (PCA), utilizando la función `prcomp` del paquete de estadística computacional R.¹⁹ El programa ADMIXTURE se ejecutó sobre el conjunto de datos de genotipo para inferir los componentes ancestrales individuales.²⁰ Los datos resultantes se utilizaron en la función `nnls` del paquete R para implementar un método de mínimos cuadrados no negativo para estimar las fracciones de ancestralidad africana, nativa americana y europea para cada individuo de la población del Chocó. Para cada individuo, la entropía (H) de la mezcla se calculó como $= \sum_{i=1}^3 -p_i * \log(p_i)$, donde p_i es la fracción de ancestralidad para la población i . La ancestralidad específica de sexo en la población del Chocó fue determinada mediante el análisis de

¹⁹ Human Genome Diversity Project

haplotipos uniparentales en el mtADN para el linaje materno y en Y-ADN para el linaje paterno.

Diversidad genética y heterocigosidad: Los niveles relativos de diversidad genética de las poblaciones analizadas aquí, se midieron a través de la cantidad total de distancias compartidas entre alelos apareados y la heterocigosidad en todo el genoma. La cantidad total de distancias compartidas entre alelos apareados dentro de cada población se calculó mediante el ajuste de una elipse de expansión mínima para los puntos de genotipos individuales de cada población proyectada sobre los dos primeros componentes principales del análisis PCA, utilizando la función *ellipsoidhull* en el paquete R. Las áreas (A) de las elipses específicas de cada población se calcularon usando las longitudes de los ejes mayores (x) y menores (y), escalados a los pesos de los componentes principales: $A = \pi(0.665 * x)(0.060 * y)$. La heterocigosidad se midió como la fracción de todas las posiciones de genotipo que son heterocigotos dentro de un individuo, utilizando el programa PLINK. Para ello, se analizaron SNPs con frecuencias de alelo menor de $\geq 25\%$ con el fin de permitir la comparación entre los SNPs de obtenidos de secuencias genómicas SNPs frente a los obtenidos a partir de genotipos. Estos últimos están sesgados hacia frecuencias altas de alelos menores y hacia poblaciones europeas.

Aspectos éticos: Todos los donantes firmaron un consentimiento informado que indica su comprensión de los riesgos potenciales del proyecto junto con la forma en que sus datos serían manejados y cómo sería protegida su identidad. La genotipificación y el análisis comparativo de las muestras de ADN humano se llevaron a cabo con el aval del comité de ética de la Universidad Tecnológica del Chocó. Una gran cantidad de datos analizados se obtuvieron a partir de bases de datos públicas en las cuales es imposible establecer conexión entre las secuencias genómicas y/o los genotipos y los donantes del ADN humano.

Resultados

Poblaciones humanas analizadas en este estudio: En total se estudiaron 912 genomas, de los cuales el 11,1 % (n=101) eran genotipificados exclusivamente para este estudio. El resto fueron obtenidos a partir de bases de datos de acceso público: 1000G y HGDP. Es de anotar que el 61,4% (n=560) de las secuencias obtenidas pertenecen a poblaciones ancestrales no mestizadas genómicamente y el 27,5% (n=251) a poblaciones americanas mestizadas. La tabla 4.1 describe las poblaciones estudiadas y los códigos de color de cada población que se usan de aquí en adelante, entre otros detalles. Las poblaciones se organizan en función de que representan las poblaciones de ascendencia no mezclados procedentes de África, Europa y las Américas o poblaciones americanas mezclados desde el Caribe, EEUU y Colombia. El Chocó se refiere a la muestra de población estudiada aquí.

Ancestralidad y mestizaje genómico en el Chocó: La población colombiana tiene una mezcla de ancestralidad genética de África, Europa y poblaciones indígenas de América

(nativos americanos), debido a los patrones históricos de la conquista y colonización del nuevo mundo.^{21,22} Por lo tanto, los genotipos del Chocó se compararon con genotipos caracterizados a partir de individuos muestreados en poblaciones representativas de estas regiones (ver Tabla 4.1) con el fin de inferir su ancestralidad y mestizaje genómico en general.

Tabla 4.1: Poblaciones analizadas en este estudio				
Grupo de datos	Color	Acrónimo	Descripción	n
Poblaciones ancestrales no mestizadas				
1000G África (n=292)		LWK	Luhya en Webuye, Kenia	99
		MSL	Mende en Sierra Leone	85
		YRI	Yoruba en Ibadan, Nigeria	108
1000G Europa (n=198)		GBR	British en Inglaterra y Escocia	91
		IBS	Población Ibérica en España	107
HGDP Nativo Americano (n=70)		KRT	Karitiana en Brasil	24
		PMA	Pima en México	25
		SUR	Surui en Brasil	21
Poblaciones americanas mestizadas				
1000G American Mestizados (n=251)		ACB	Afrocaribeños en Barbados	96
		ASW	Afroamericanos en EEUU	61
		CLM	Colombianos de Medellín, Colombia	94
El Chocó (n=101)		CHO	Colombianos del Chocó, Quibdó, Colombia	101

Las relaciones genómicas entre individuos de la población del Chocó con individuos de otras poblaciones mundiales se representan visualmente en la figura 4.2A. Este panel muestra una proyección bidimensional del análisis PCA sobre las distancias genéticas entre todos los genotipos analizados aquí, donde la distancia entre cada punto corresponde a la distancia entre cada genotipo individual. El principal componente de la diversidad genética humana en esta representación se proyecta a lo largo del eje x (PC1=66,5% de la diversidad) y el componente secundario se muestra en el eje y (PC2=6,0%). África, Europa y las poblaciones nativas americanas ocupan los tres polos de la diversidad genómica humana en la figura 4.2A, mientras que los genomas mezclados americanos, incluidos los individuos Chocó (CHO), ocupan posiciones intermedias entre estos tres grupos de poblaciones ancestrales, relativamente no mestizadas. Las posiciones relativas de las poblaciones mestizadas de América en comparación con los tres grupos ancestrales dan una indicación de sus proporciones de mezcla. Por ejemplo, la población colombiana de Medellín (CLM) muestra evidencia de más proporción de mezcla europea en comparación con la población del Chocó, que se encuentra mucho más próxima a las poblaciones africanas. Otras dos poblaciones de afrodescendientes del nuevo mundo (ACB y ASW) también se encuentran en las proximidades de las poblaciones ancestrales putativos del continente africano, pero ocupan diferentes posiciones que la población del Chocó.

Estas mismas distancias genéticas apareadas se pueden utilizar para cuantificar la cantidad de ancestralidad genómica que cualquier individuo mestizado contiene a partir

de las poblaciones ancestrales putativas de África, Europa y nativos americanos. Los resultados de este tipo de análisis se muestran en la Figura 4.2B y 4.2C.

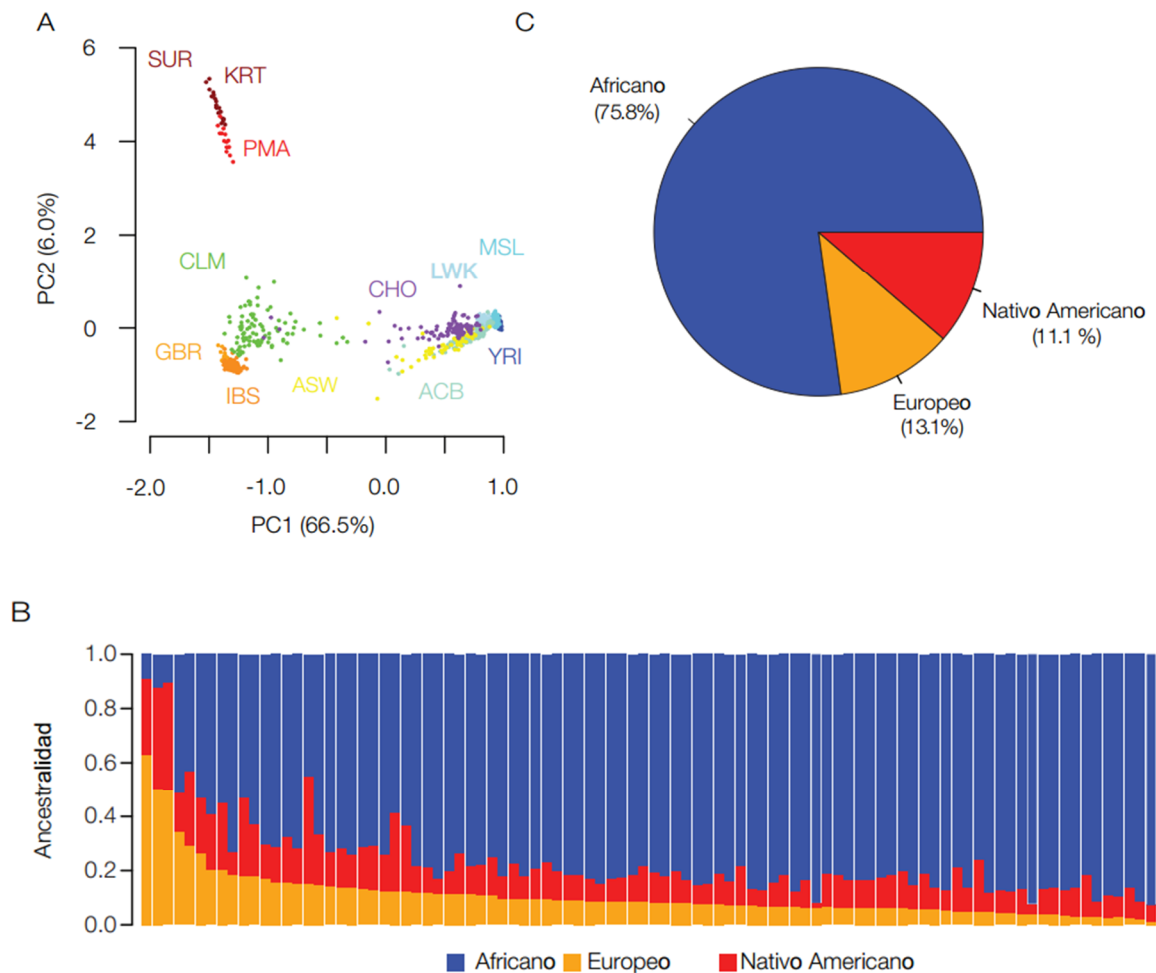


Figure 4.2: Ancestralidad y mestizaje genómico en el Chocó. A) Proyección del análisis de PCA mostrando las relaciones genómicas entre individuos de diferentes poblaciones humanas comparados con los del Chocó. Cada punto corresponde a un único individuo y las distancias entre los puntos corresponde a distancias genómicas entre los individuos. Los códigos de cada color se muestran en la figura 1.1, los del Chocó se muestran en purpura y se les asignó el acrónimo CHO. Los componentes principales se denotan como PC1 y PC2, con la cantidad de variación genética para cada componente entre paréntesis. B) Gráficos de barras que muestra el mestizaje genómico de cada individuo de la población del Chocó a partir de las tres poblaciones ancestrales: África (azul), Europa (naranja) y Nativo Americano (rojo). C) Pastel que representa los valores promedios de ancestralidad para toda la muestra del Chocó. Los códigos de color son los mismos que en la figura 4.2B.

La ancestralidad africana representa el componente dominante del mestizaje para la gran mayoría de los individuos de la población del Chocó analizados aquí. La fracción máxima de ancestralidad africana observada fue del 92,8%, con un promedio de 75,8% en toda la muestra analizada del Chocó. No obstante, hay fracciones sustanciales de ancestralidad europea y nativa americana en muchos los individuos del Chocó. La fracción máxima de

ancestralidad europea observada fue del 62,8%, con un promedio de 13,1%. La fracción máxima de ancestralidad fue del 39,6%, con un promedio del 11,1%. La amplia gama de porcentajes de mestizaje individual apuntan hacia la diversidad de la población del Chocó.

La ancestralidad genómica de la población del Chocó muestra algunas diferencias interesantes en comparación con las otras dos poblaciones del afrodescendientes del nuevo mundo analizadas aquí, la población ACB de Barbados en el Caribe y la población ASW de los EEUU, a pesar del hecho de que las tres poblaciones muestran igualmente altos niveles de ancestralidad africana en general (~75-80%). En primer lugar, la población del Chocó (CHO) tiene niveles sustancialmente más altos de ancestralidad nativa americana en comparación con las poblaciones afroamericanas (ASW) o afrocaribeña (ACB) (Figura 4.3A). Chocó tiene un promedio de 11,1% de ancestralidad nativa americana, mientras que las poblaciones afrocaribeñas y afroamericanas tienen un 0,4% y 1,1%, respectivamente. En segundo lugar, la población del Chocó muestra los niveles más altos de mestizaje genómico triple, según la entropía de mestizaje (H) determinada, en comparación con las otras dos poblaciones afrodescendientes del nuevo mundo (Figura 4.3B). Esto refleja el hecho de que en el Chocó el componente de ancestralidad no africana es relativamente uniformemente distribuida entre las ancestralidades europea y nativa, mientras que casi toda la ancestralidad no africana en las poblaciones del Caribe y Estados Unidos es europea. Este patrón es indicativo de contacto más largo y más sostenido entre los afrodescendientes y las comunidades indígenas del Chocó en comparación con lo ocurrido en el Caribe o en los Estados Unidos.

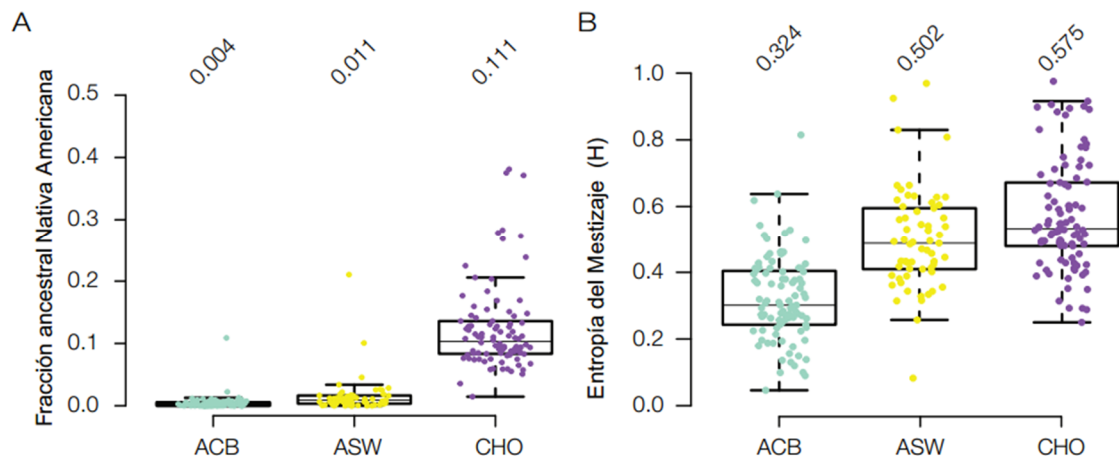


Figura 4.3: Características distintivas del mestizaje en el Chocó. Se comparan los patrones de mestizaje entre Chocó (CHO) y otras poblaciones del afrodescendientes del nuevo mundo: ACB de Barbados en el Caribe y ASW de los Estados Unidos. A) Ancestralidad nativa americana y (B) Distribuciones del mestizaje genómico en las tres poblaciones afrodescendientes del nuevo mundo. Los valores promedio se muestran para cada distribución.

Ancestralidad genómica específica de sexo: El ADN mitocondrial (mtADN) y el cromosoma Y (Y-DNA) se denominan marcadores de ancestralidad uniparental dado que se heredan

estrictamente por linaje materno (mtADN) y paterno (Y-ADN), respectivamente. Esto significa que los haplotipos de mtADN pueden utilizarse para inferir ancestralidad y mestizaje de origen femenino, mientras que los haplotipos Y-ADN se pueden usar para inferir ancestralidad y mestizaje de origen masculino. Los orígenes globales de los haplotipos del mtADN y Y-ADN caracterizados a partir de la población del Chocó muestran evidencia notable de ancestralidad sexo específica en esta población (ver Figura 4.4). La mayoría de los haplotipos del mtADN (82,1%) y del Y-ADN (77,8%) tienen orígenes africanos, consistente con la ancestralidad genómica de la población general. Sin embargo, los componentes de ancestralidad no africana difieren notablemente entre el linaje materno y el paterno. Todos los haplotipos de mtADN no africanos (17,9%) tienen un origen nativo americano, mientras que todos los haplotipos Y-DNA no africanos tienen de orígenes europeo (16,7%) o de Oriente Medio (5,6%). Este patrón específico del sexo de la ascendencia genética puede estar vinculada a las condiciones históricas únicas en virtud de los cuales se fundó el estado de Chocó y pobladas.²³

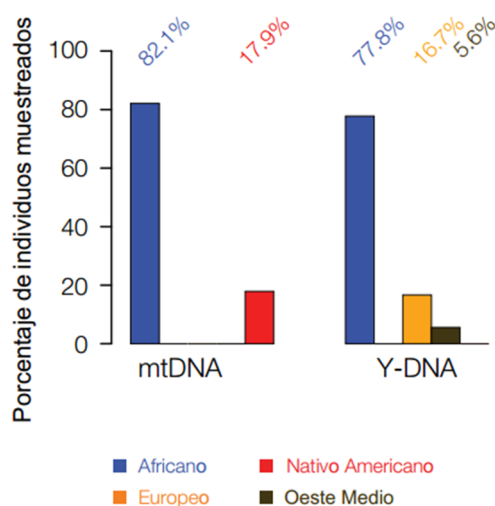


Figura 4.4: Ancestralidad y mestizaje específico de sexo en El Chocó. Se muestran las proporciones de haplotipos de varios orígenes: África (azul), Nativo Americano (rojo), Europa (naranja) y Oriente Medio (marrón) del mtADN (linaje materno) y Y-ADN (linaje paterno).

Niveles de diversidad genética en el Chocó: Como se señaló anteriormente, con base en su demografía poblacional distintiva, se propone que el departamento del Chocó es una rica fuente de biodiversidad humana. Para evaluar esta propuesta con respecto a la ancestralidad genómica de la población, se compararon los niveles de diversidad genética que se encuentran en Chocó con los niveles de diversidad visto en poblaciones ancestrales putativas, no mestizadas; así como con otras poblaciones americanas mestizadas. Los resultados de este análisis se muestran en la Figura 4.5. Se evaluó la diversidad genética de dos maneras: 1) A través del alcance general de las distancias genéticas entre

individuos de una población y 2) a través de los niveles de heterocigosidad genómica promedio para todos los individuos en una población.

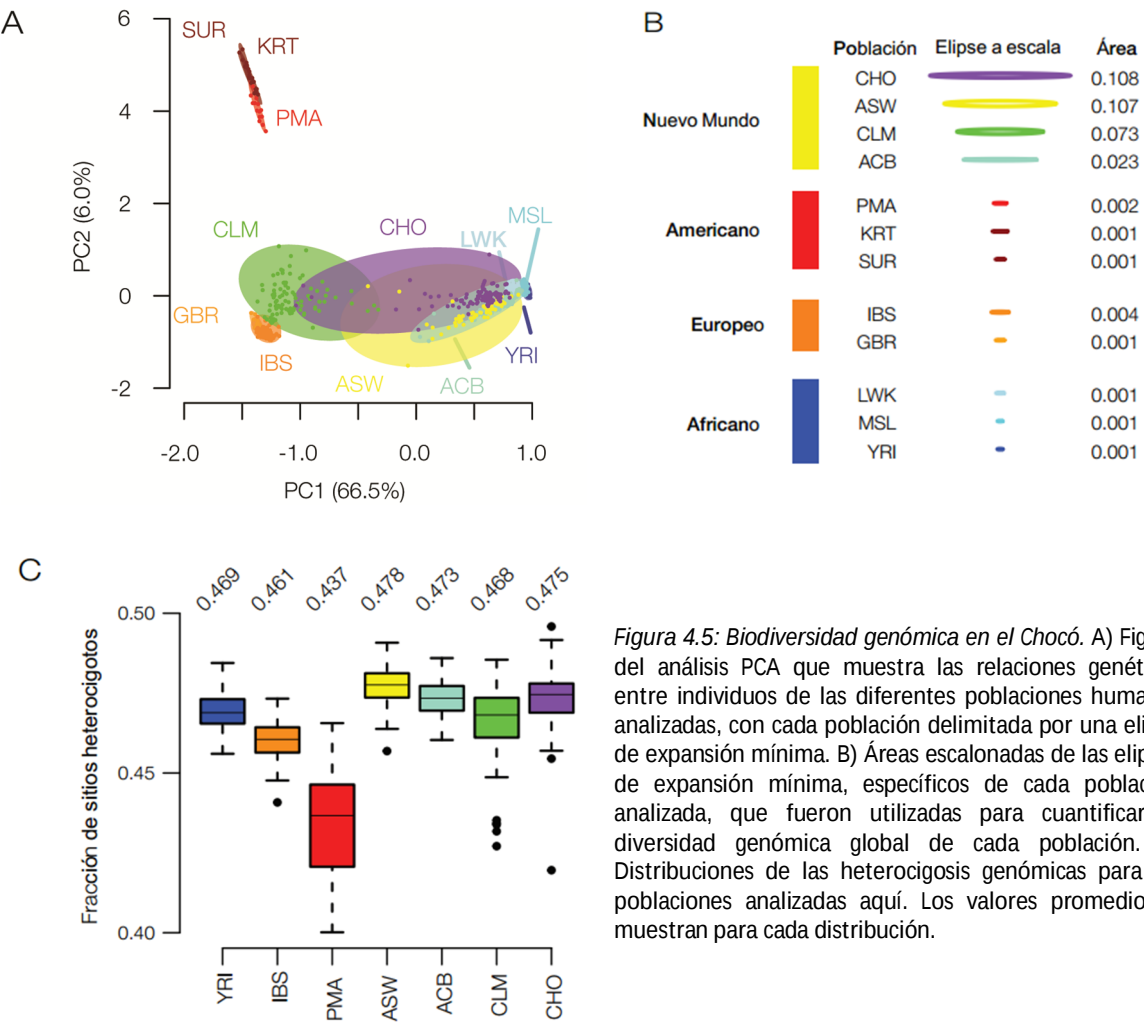


Figura 4.5: Biodiversidad genómica en el Chocó. A) Figura del análisis PCA que muestra las relaciones genéticas entre individuos de las diferentes poblaciones humanas analizadas, con cada población delimitada por una elipse de expansión mínima. B) Áreas escalonadas de las elipses de expansión mínima, específicos de cada población analizada, que fueron utilizadas para cuantificar la diversidad genómica global de cada población. C) Distribuciones de las heterocigosidades genómicas para las poblaciones analizadas aquí. Los valores promedio se muestran para cada distribución.

La diversidad genética global para cada población se infirió por el ajuste de una elipse de expansión mínima a la proyección de las distancias genéticas apareadas de las poblaciones en la figura del análisis PCA (Figura 4.5A). Las áreas de las elipses de expansión mínima en cada población específica fueron entonces utilizadas para cuantificar la diversidad genética de las poblaciones (Figura 4.5B). La población del Chocó (CHO) tiene el más alto nivel de diversidad genética total calculado de esta manera para cualquiera de las poblaciones analizadas aquí. La población del Chocó también tiene altos niveles de heterocigosidad promedio en comparación con los otros grupos poblacionales, siendo sólo superada por la población ASW, una de las afrodescendientes del nuevo mundo (Figura 4.5C). Curiosamente, la otra población colombiana mestizada de Medellín (CLM)

también mostró niveles relativamente altos de diversidad genómica en estos análisis a pesar del hecho de que tiene un perfil muy diferente ancestralidad, es decir predominantemente europea, en comparación con la población del Chocó.

Discusión

Los análisis aquí presentados indican que la población del Chocó tiene una ancestralidad genómica predominantemente africana, lo cual por supuesto no es una sorpresa. Sin embargo, la población del Chocó muestra diferencias interesantes, en comparación con otras poblaciones mestizadas americanas con niveles similares de ancestralidad africana, lo que probablemente refleja sus diferentes tradiciones históricas y culturales.²³ En particular, los individuos de Chocó muestran niveles más altos de mestizaje genómico triple que otras poblaciones afrodescendientes del nuevo mundo, y este patrón se puede atribuir en gran parte a los niveles más altos de ancestralidad nativa americana observadas en El Chocó. La población del Chocó también muestra patrones muy interesantes de ancestralidad específica de sexo, evidenciados por la [proporción de ancestralidad no africana que es exclusivamente de nativos americanos en el linaje materno y casi en su totalidad europea en el linaje paterno. Este patrón de ancestralidad puede representar las dinámicas de mestizaje que diferenciaron los periodos históricos tempranos (mestizaje europeo) de los posteriores (mestizaje nativo americanos) en el Chocó.

Los resultados en la ancestralidad genómica y la diversidad genética obtenidos en este estudio subrayan el grado en que el Chocó representa un *hotspot* de la biodiversidad humana. Es posible entonces asegurar que la biodiversidad humana del Chocó es una dimensión poco apreciada de la biodiversidad ya conocida de la zona y que debe ser igualmente valorado y completamente desarrollado para evidenciar todo su potencial.

Este capítulo de la tesis hace parte de un macroproyecto de investigación colaborativa internacional, denominado ChocoGen²⁰ cuyos objetivos son 1) Caracterizar la ancestralidad genómica y la diversidad genética de la población del Chocó y 2) Crear perfil genético de salud/enfermedad de la población basado en la diversidad de su gente.

Referencias capítulo 4

1. Zachos, F.E. & Habel, J.C. *Biodiversity hotspots: distribution and protection of conservation priority areas*, (Springer Science & Business Media, Vienna, 2011).
2. Hernández Romero, A. *La visibilización estadística de los grupos étnicos colombianos*, (Departamento Administrativo Nacional de Estadística (DANE), Bogotá, 2005).
3. Rishishwar, L., Conley, A.B., Vidakovic, B. & Jordan, I.K. A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians. *Gene* 574, 345-351 (2015).
4. Bedoya, G., et al. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7234-7239 (2006).

²⁰ <http://www.chocogen.com/>

5. Bryc, K., *et al.* Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* 107 Suppl 2, 8954-8961 (2010).
6. Carvajal-Carmona, L.G., *et al.* Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Human genetics* 112, 534-541 (2003).
7. Carvajal-Carmona, L.G., *et al.* Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *American journal of human genetics* 67, 1287-1295 (2000).
8. Cordoba, L., *et al.* Composicion genetica de una poblacion del suroccidente de Colombia. *Revista Colombiana de Antropologia* 48, 21-48 (2012).
9. Rishishwar, L., *et al.* Ancestry, admixture and fitness in Colombian genomes. *Scientific reports* 5, 12376 (2015).
10. Ruiz-Linares, A., *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS genetics* 10, e1004572 (2014).
11. Wang, S., *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS genetics* 4, e1000037 (2008).
12. Genomes Project, C., *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073 (2010).
13. Genomes Project, C., *et al.* A global reference for human genetic variation. *Nature* 526, 68-74 (2015).
14. Cann, H.M., *et al.* A human genome diversity cell line panel. *Science* 296, 261-262 (2002).
15. Li, J.Z., *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100-1104 (2008).
16. Kent, W.J., *et al.* The human genome browser at UCSC. *Genome research* 12, 996-1006 (2002).
17. Lander, E.S., *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001).
18. Purcell, S., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81, 559-575 (2007).
19. Team, R.D.C. *R: A language and environment for statistical computing.*, (R Foundation for Statistical Computing, Vienna, 2008).
20. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19, 1655-1664 (2009).
21. Mann, C.C. *1493: Uncovering the new world Columbus created*, (Alfred a Knopf, New York, 2013).
22. Markham, C. *The Conquest of New Granada*, (EP Dutton and Company, New York, 1912).
23. Wade, P. *Blackness and race mixture: the dynamics of racial identity in Colombia*, (JHU Press, Balitmore, 1995).

Conclusiones

- Los hallazgos permiten subrayar la importancia del papel de la introgresión como fuente de alelos de adaptación y como motor de cambio evolutivo; además, pone de relieve el papel del mestizaje como facilitador de la rápida evolución humana.
- El mestizaje genómico de los colombianos es el de mayor extensión en los países de Latinoamérica con contribución de tres poblaciones ancestrales: europeos, africanos y nativos americanos; con una clara divergencia entre la autoidentificación étnica y la ancestralidad genética.
- Estos resultados sugieren la posibilidad de que algunos *loci* de ancestralidad específica han sido retenidos de forma diferencial en la población colombiana moderna, genómicamente mezclada, en función de su utilidad en el entorno del llamado Nuevo Mundo; en particular aquellos relacionados con la respuesta inmune.
- Los resultados con la población del Chocó sugieren un patrimonio genético único que ponen de relieve la profunda diversidad genética humana que se pueden encontrar en la región.
- En las dos poblaciones analizadas, el mestizaje genético específico de sexo demuestra claramente que la ancestralidad europea fue dominante en el linaje paterno, mientras que la nativa americana fue dominante en el linaje materno.

Recomendaciones

Los resultados de esta tesis de doctorado muestran una conexión explícita entre la ancestralidad genética humana y los determinantes genéticos de la salud y la adaptación al medio. La relación entre la ancestralidad y los determinantes genéticos de la salud humana, a menudo se manifiestan como disparidades del perfil salud/enfermedad de las poblaciones, es por tanto un tema importante con consecuencias graves para la salud pública. Por ejemplo, la investigación sobre cómo las poblaciones mestizadas han sido moldeadas por las presiones selectivas impuestas por la carga de enfermedades infecciosas puede dar una idea de la arquitectura genética de la respuesta inmune. El estudio de los genomas mestizados a través de todas las Américas podría proporcionar una importante extensión a los estudios genómicos clínicos actuales, de los cuales la gran mayoría se han concentrado en las poblaciones de ancestralidad europea.

Una recomendación lógica y evidente es por tanto, continuar estas investigaciones, hacia la creación de un perfil genómico específico y característico de las poblaciones que habitan actualmente el territorio colombiano.

Anexos

Anexo 1: Manuscrito publicado en *Nature Scientific Reviews*.

Anexo 2: Manuscrito publicado en *Revista de Biodiversidad Neotropical*

SCIENTIFIC REPORTS

OPEN

Ancestry, admixture and fitness in Colombian genomes

Lavanya Rishishwar^{1,2,3,*}, Andrew B. Conley^{1,*}, Charles H. Wigington¹, Lu Wang¹, Augusto Valderrama-Aguirre^{2,4,5} & I. King Jordan^{1,2,3}

Received: 31 March 2015

Accepted: 01 June 2015

Published: 21 July 2015

The human dimension of the Columbian Exchange entailed substantial genetic admixture between ancestral source populations from Africa, the Americas and Europe, which had evolved separately for many thousands of years. We sought to address the implications of the creation of admixed American genomes, containing novel allelic combinations, for human health and fitness via analysis of an admixed Colombian population from Medellín. Colombian genomes from Medellín show a wide range of three-way admixture contributions from ancestral source populations. The primary ancestry component for the population is European (average = 74.6%, range = 45.0%–96.7%), followed by Native American (average = 18.1%, range = 2.1%–33.3%) and African (average = 7.3%, range = 0.2%–38.6%). Locus-specific patterns of ancestry were evaluated to search for genomic regions that are enriched across the population for particular ancestry contributions. Adaptive and innate immune system related genes and pathways are particularly over-represented among ancestry-enriched segments, including genes (*HLA-B* and *MAPK10*) that are involved in defense against endemic pathogens such as malaria. Genes that encode functions related to skin pigmentation (*SCL4A5*) and cutaneous glands (*EDAR*) are also found in regions with anomalous ancestry patterns. These results suggest the possibility that ancestry-specific loci were differentially retained in the modern admixed Colombian population based on their utility in the New World environment.

The arrival of Columbus in the New World precipitated a massive and sudden exchange of life forms between the American and Afro-Eurasian hemispheres. European colonization and trade in the Americas entailed a bidirectional exchange of numerous species of plants, animals and microbes, many of which until that time had been evolving separately for millions of years. The conquest and settlement of the Americas, along with the African slave trade, also included exchanges among a variety of human populations. These biological and cultural exchanges, which were often intentional although sometimes not so, had a tremendous impact on the trajectory of human history in the centuries that followed. Among other effects, the exchange of crops and livestock allowed for a major increase in the global population, whereas the introduction of microbes that cause infectious disease had devastating effects on the naïve populations to which they were introduced.

The historian Alfred Crosby referred to this massive transfer of life as the *Columbian Exchange* in his seminal book of the same name¹. Numerous dimensions of the Columbian Exchange have been studied since the concept was first introduced, and these ideas have received renewed attention with the recent publication of the book *1493: Uncovering the New World Columbus Created* by Charles Mann². In particular, the ecological, economic and cultural-historical aspects of this era have been explored at length. One especially intriguing aspect of the Columbian Exchange that has not been addressed within this conceptual framework is the exchange that occurred at the level of human genome sequences.

¹School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. ²PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia. ³BIOS Centro de Bioinformática y Biología Computacional, Manizales, Caldas, Colombia. ⁴Biomedical Research Institute, Universidad Libre, Cali, Valle del Cauca, Colombia. ⁵Regenerar - Center of Excellence for Regenerative and Personalized Medicine, Cali, Valle del Cauca, Colombia. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to I.K.J. (email: king.jordan@biology.gatech.edu)

Ethnic Group	Colombia (specific) ¹	Colombia (broad) ¹	Medellin (broad) ²
Mestizo (Amerindian and European)	58%	78%	93.4%
White (European)	20%		
Mulatto (Black/ African and European)	14%	21%	6.5%
Black (African)	4%		
Zambo (Amerindian and Black/ African)	3%		
Amerindian	1%	1%	0.1%

Table 1. Demographic profile of ethnic groups for Colombia and Medellin. ¹Colombian ethnic group percentages are taken from the CIA World Factbook¹⁸. Specific and broad ethnic groupings are shown. ²Medellin ethnic group percentages are taken from the Colombian census³⁶. Only broad ethnic groupings are available for individual cities.

The modern human species emerged out of Africa and spread throughout the world starting between 60–100,000 years ago³. Human evolution during this time was primarily characterized by migration and geographical isolation, followed by population divergence. Over tens-of-thousands of years, these processes gave rise to the major continental groups of human populations recognized today: African, European, Asian, Melanesian and American^{4,5}. The trajectory of human evolution was turned upside-down during the course of the Columbian Exchange. Within the last 500 years, populations that were separated for many thousands of years were brought back together, and as tends to be the case whenever humans are placed in close proximity, they then began to exchange genes. This process of genetic admixture in the Americas has occurred over an extremely short time period in human evolution, for <1% of the time since modern humans first emerged from the African continent.

Admixture rapidly brings together population-specific (or enriched) alleles that have not previously co-existed in the same genetic background, and thereby can be considered to result in the creation of completely novel human genomes. Population-specific (enriched) alleles are sequence variants that evolved to population-characteristic frequencies *in situ* within ancestral populations' endemic geographic regions. Some of these alleles may have drifted to high frequency within the ancestral populations by chance, whereas others are likely to have been swept to fixation based on selection pressures that were distinct to the ancestral populations' environments^{6–9}. Genes that mediate humans' interaction with their environment, such as those that encode skin pigmentation and immune system related proteins, seem to have been particularly prone to adaptive evolution in ancestral regions. For example, there are several skin pigmentation genes with European-specific and/or Asian-specific alleles that are associated with lighter skin color^{6–8,10–16}. A number of genes involved in the defense against infectious pathogens also evolved regional-specific alleles that are concordant with the ranges of particular pathogens^{6–10}. At least three distinct genes have evolved population-specific alleles related to the defense against malaria in regions endemic for the disease^{7,8,10}. In general, populations from pathogen-rich global regions, including regions of West Africa, East Asia and the Americas, encode a more diverse repertoire of immune receptors, and this is thought to be due to selection pressure to confront a wider variety of microbial pathogens¹⁷.

In this study, we sought to address what it means when genomic variants that have been separated for tens-of-thousands of years are suddenly brought back together over the course of a few hundred years. In other words, is the process of modern admixture somehow related to human health and fitness? Our working hypothesis is based on the well-supported axiom that specific allelic variants (SNPs) have evolved separately in ancestral human populations based on their regional-specific utility, *i.e.* their relationship to health and fitness in a particular environment. We posit that these pre-evolved ancestral population-specific alleles may have been selected in the modern admixed population based on their utility in the new environment. The new environment could be a new physical environment, for populations that were transported to new regions, and/or a new ecological environment based on the mixing of previously isolated human populations and their associated microbial fauna.

We evaluate these ideas here via a study of the relationship between ancestry, admixture and fitness in Colombian genomes. The Colombian population is a particularly interesting subject for study in this regard owing to its high levels of ethnic admixture¹⁸. The people of Colombia are richly diverse with substantial admixture between African, Native American and European ancestral populations (Table 1)^{19–23}. In fact, it has been reported that Colombians show among the greatest extent of three-way continental admixture of all genetically characterized Latino/Hispanic populations^{24,25}. We analyzed the ancestry and admixture patterns for whole genome sequences of 60 unrelated Colombians from Medellin, which were recently sequenced as part of the 1000 Genomes Project^{26,27}. We then developed and applied a method to search for genomic regions that show anomalous patterns of admixture based on differential retainment

of chromosomal segments from specific ancestral populations in the modern admixed population. These anomalous regions were interrogated for previously identified signatures of natural selection as well as for the functional and health-associated roles of the genes encoded therein. This approach yielded results indicating that there is substantial enrichment for ancestry-specific loci genome-wide, and these regions encode numerous genes involved in immune related functions along with genes previously implicated in selection for skin pigmentation and glandular development.

Results

Ancestry and admixture in Colombian genomes. Admixed Colombian genome sequences from Medellin were compared with a number of world-wide populations that are likely to be closely related to Colombian ancestral source populations (Supplementary Fig. S1). To do this, pairwise allele (SNP) sharing distances were computed between all genomes and principal component analysis (PCA) was used to project the resulting pairwise distances (Fig. 1A and Supplementary Fig. S2). The first principal component (81.1% of the variation) shows clear separation between the African ancestral population and all other populations, whereas the Native American, East Asian and European ancestral populations are separated along the second principal component (10.5% of the variation). The Colombian genomes from Medellin appear most closely related to the European ancestral population, but extend outward along both principal component axes consistent with African and Native American admixture. The East Asian population is most closely related to the Totonac Native American population from Mexico followed by the Bolivian population, which shows higher levels of apparent European admixture.

The close clustering of the East Asian with the Bolivian Native American populations (Fig. 1A) is consistent with the relatively recent origin of Native Americans from Asia²⁸ and suggests that Asian genomes are a viable surrogate for inferring Native American ancestry contributions at the continental level. In other words, when admixed Colombian genomes are analyzed in three-way continental level comparisons to ancestral source populations, including African and European populations, East Asian genomes are very likely to recover Native American ancestral genomic segments. Whole East Asian (CHB) genome sequences were used here, along with whole African (YRI) and European (CEU) genome sequences, for admixture analysis of the Colombian genomes in an effort to provide additional resolution beyond what is available from relatively sparse genotype data. An admixture plot showing the three ancestral genome clusters, together with the admixed Colombian genome cluster, is shown in Fig. 1B along with a plot expanding the Colombian genomes (Fig. 1C). The Colombian genomes show substantial variability in admixture patterns with different levels of contribution from ancestral populations. The percent ancestry values for individual genomes range from 1.2% African, 2.1% Asian (Native American) and 96.7% European to 36.8% African, 18.2% Asian (Native American) and 45.0% European. The admixture contributions to the Colombian genomes from ancestral populations are well outside the error levels seen based on the extent of false-positive ancestry assignment levels in the ancestral clusters (Fig. 1B). The average ancestry values for the admixed Colombian genomes from Medellin are 7.3% African, 18.1% Asian (Native American) and 74.6% European (Fig. 1D).

Sex-specific admixture. A number of previous studies have uncovered sexual asymmetry in the contributions of different ancestral populations to admixed Latino genomes^{20–23,25}. Admixed Latino genomes tend to show a relative excess of European paternal ancestry and proportionally greater Native American maternal ancestry. We evaluated the sexual asymmetry of admixture in the Colombian population by comparing ancestry contributions to the X chromosome versus the autosomes as described in the Materials and Methods. Since X chromosomes spend relatively more time along the female lineage, a relative excess of a specific ancestral admixture component in X chromosomes indicates a proportionally greater female (maternal) contribution for that ancestry. Conversely, a relative excess of a specific ancestral admixture component in autosomes indicates a proportionally greater male (paternal) contribution. Consistent with previous reports, the Colombian population shows highly sex-specific admixture patterns with predominantly European contributions to the male lineage and Native American ancestry along the female lineage (Fig. 2). Interestingly, this pattern is more pronounced for the Colombian population than for any of the other four Latino populations with which it was compared. The relative level of European male ancestry for the Colombian population is significantly greater than seen for that of the country with the next highest level (Puerto Rico; $P = 1.8 \times 10^{-7}$). Similarly, the relative level of Native American female ancestry was significantly higher for the Colombian population compared with the country with the next highest level (Ecuador; $P = 1.2 \times 10^{-7}$). The Colombian population also shows a smaller, but not insubstantial, excess of African ancestry along the male lineage (African paternal ancestry differs from 0 at $P = 3.2 \times 10^{-14}$).

Population-wide admixture enrichment. The three-way continental ancestry origins of individual chromosomal segments across the entire genome were determined for all 60 individuals in the Colombian population as described in the Materials and Methods. Chromosome paintings that show the locations of ancestry-specific segments across the chromosomes of two Colombian individuals are shown in Fig. 3; similar results for all 60 Colombian genomes analyzed here can be found in the Supplementary Video. As can be seen from these plots, along with Fig. 1B–D, individual Colombians vary substantially with respect to both their overall three-way continental ancestry contributions and their regional (locus-specific)

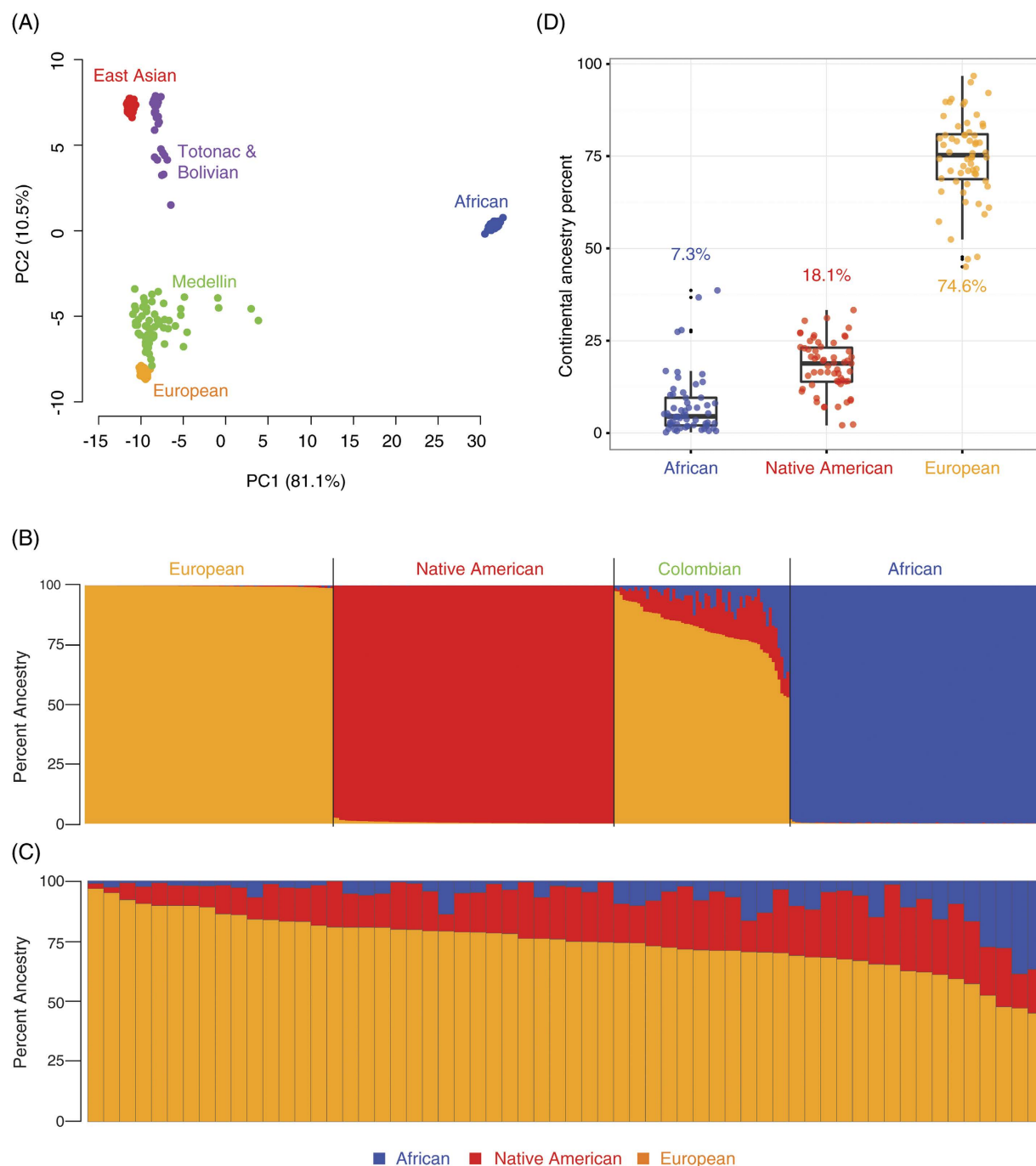


Figure 1. Ancestry and admixture patterns for Colombian genomes. (A) PCA of pairwise allele sharing distances among admixed Colombian genomes from Medellin compared with putative ancestral populations from Africa, the Americas, Asia and Europe. (B) Admixture plots showing ancestry proportions for three putative ancestral populations and the admixed Colombian genomes. (C) The lower panel shows only the Colombian genomes. (D) Ancestral admixture proportion distributions and averages for the Colombian genomes.

ancestry origins. The locus-specific patterns of ancestry among the Colombian population were analyzed to search for anomalous genomic regions that are enriched for contributions from a particular ancestral population. To do this, the population-wide ancestry profiles of individual chromosomal segments were compared with the overall average ancestry contributions as described in the Materials and Methods. The boundaries of the individual chromosomal segments (loci) used for this ancestry enrichment analysis are defined by recombination maps as previously described²⁹. There are 379,218 such loci genome-wide and the average length of an ancestry block is 7,542 bp (see loci size distribution in Supplementary Fig. S3).

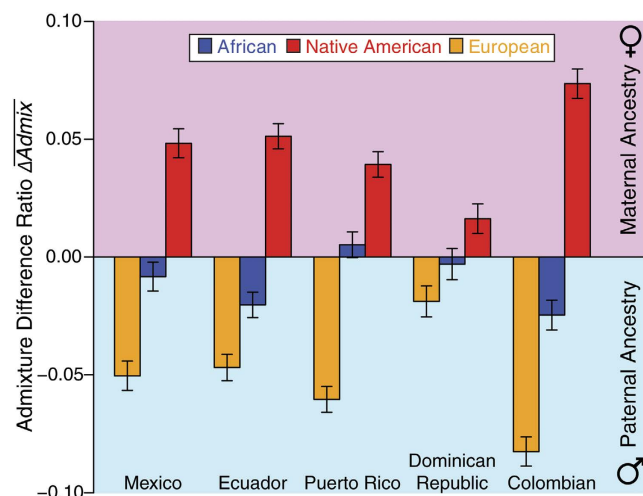


Figure 2. Sex-specific contributions to three-way genomic admixture in Latin American countries. For each ancestry component – African, Asian (Native American) and European – the normalized difference between the X chromosome ancestry fraction and the autosomal ancestry fraction is shown. Positive values in the plot indicate a relative excess of female-specific (maternal) ancestry for a given admixture component, whereas negative values indicate an excess of male-specific (paternal) ancestry.

The rationale of the ancestry enrichment approach is illustrated in Fig. 4A; ancestry-enriched segments are identified by virtue of having anomalously high levels of a specific ancestry contribution compared with the expected values based on population-wide average ancestry proportions. The frequencies of the different three-way ancestry proportion combinations in the Colombian population are represented as a heatmap in Fig. 4B. More common ancestry proportion combinations are shown as hot (red) regions, whereas the cold (blue) regions show less likely (*i.e.* anomalous) ancestry proportion combination values. Chromosomal segments that bear such anomalous ancestry proportion combinations are the ones that are identified by the ancestry-enrichment statistical test applied here (see Materials and Methods). When the ancestry-enrichment analysis technique is applied to population-wide ancestry proportion combination data, numerous statistically significant ancestry-enriched chromosomal segments are revealed across the entire genome (Fig. 4C). A list of all ancestry-enriched chromosomal segments, along with the genes that lie therein, is provided as Supplementary Table S1.

Health and selection related genes in ancestry-enriched regions. Having identified genomic loci with anomalous ancestry contribution levels, we then interrogated the genes located in these regions for previously identified signatures of natural selection and for associations with health-related traits. Only ancestry-enriched chromosomal segments that were highly statistically significant ($P < 10^{-9}$, FDR q-value 6.5×10^{-9}) were used for these analyses. There are numerous genes in these regions that were previously identified to show evidence of positive selection in populations from one of the three ancestral regions that contributed to the modern admixed Colombian population; there are also a number of genes in ancestry-enriched regions that show evidence for a role in various health-related phenotypes based on previous selection and association studies (Table 2). In particular, there are several immune system related genes that have been previously characterized as subject to positive selection found in ancestry-enriched regions (*CD226*, *HLA-B*, *MICA* and *MAPK10*).

The *HLA-B* gene located in the major histocompatibility complex (MHC) on chromosome 6 is found in an African ancestry-enriched chromosomal segment (Fig. 5). *HLA-B* is an MHC class I gene, which encodes a cell surface protein that presents foreign peptides (antigens) to immune system cells. *HLA-B* is a highly diverse gene with numerous allelic variants; high sequence diversity of *HLA* genes is thought to facilitate the ability to counter a diverse repertoire of foreign antigens. Indeed, *HLA* gene diversity levels among global populations are positively correlated with regional-specific levels of pathogen richness, with African populations having the most diverse repertoire of *HLA* alleles¹⁷. In addition, *HLA-B* allelic diversity levels world-wide show a strong positive correlation with malaria selective pressure³⁰. Enrichment for African alleles of the *HLA-B* locus may have helped the Colombian population to more effectively counter malaria and/or other pathogenic agents. The *MAPK10* gene is another immune-related gene that is located in an Asian-enriched ancestry segment in the Colombian population, and alleles of this gene were recently shown to be selected for malaria resistance in Malaysia³¹.

Some other interesting genes from ancestry-enriched regions include *SCL4A5* and *EDAR*. The *SCL4A5* gene is located in a genomic region enriched for European ancestry in the Colombian population. Positive selection at this locus has previously been associated with the evolution of decreased melanin skin pigmentation as European and Asian populations radiated out of Africa¹². Light skin has been

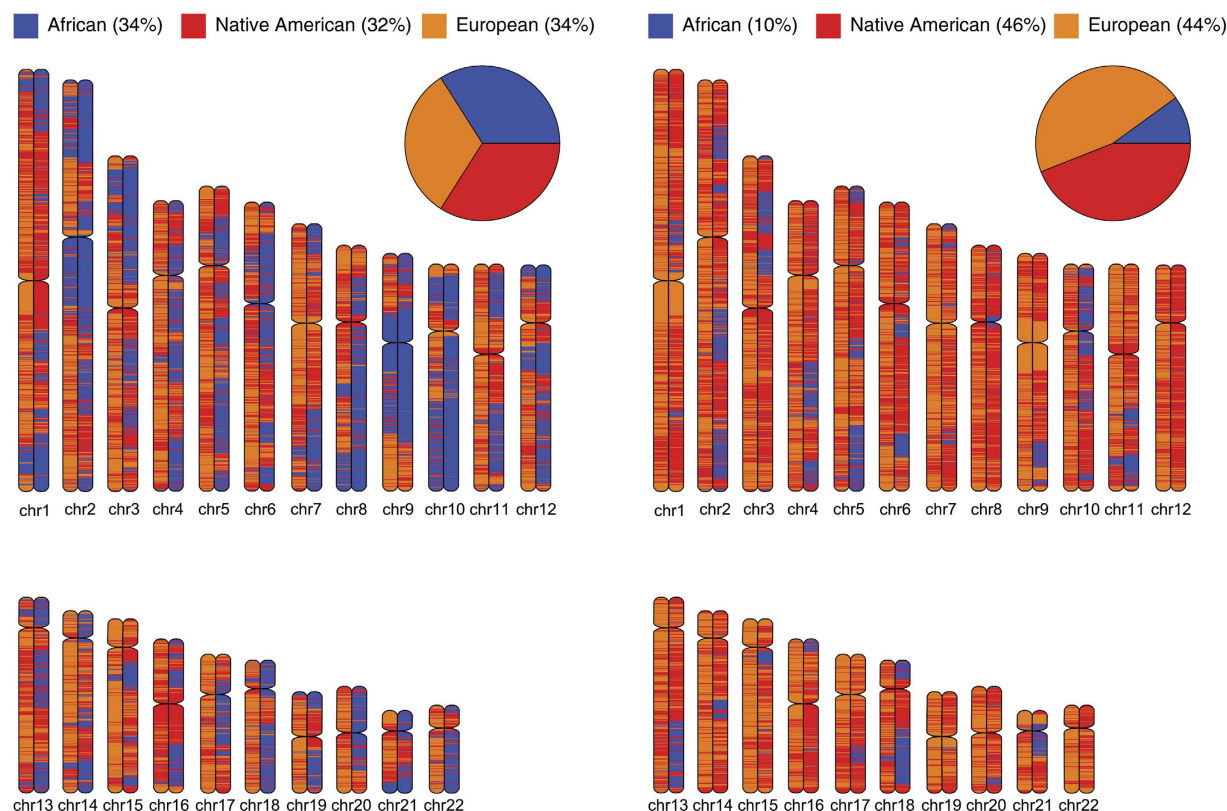


Figure 3. Regional (locus-specific) ancestry and admixture in Colombian genomes. Chromosome paintings showing the genomic distributions of loci with African, Asian (Native American) and European ancestry, along with their genome-wide ancestry proportions, are shown for two example Colombian individuals.

proposed to have evolved via sexual selection based on male preferences for lighter-skinned mates³², as originally proposed by Darwin, although this hypothesis has been contested³³. It is interesting to speculate as to whether enrichment for European ancestry at this locus can be attributed to preference for lighter-skinned females in Colombia.

The *EDAR* gene has received considerable attention recently owing to an interesting connection between the phenotype it encodes and evidence of population-specific positive selection^{9,34}. *EDAR* encodes a cell surface receptor that is involved in the development of hair follicles and cutaneous glands, and there is evidence for positive selection of *EDAR* alleles in Asian and Native American populations. The selected allele reduces evaporation from exposed facial structures and upper airways, which is thought to represent an adaptation to cold and dry environments in East Asia. In the admixed Colombian population, *EDAR* is found in a chromosomal segment that is depleted for Asian (Native American) ancestry (Supplementary Fig. S4). This could reflect the fact that the *EDAR* adaptation to cold and dry environments would confer a disadvantage in the hotter and more humid tropical environment of Colombia.

Immune related pathway genes in ancestry-enriched regions. In addition to the targeted analysis approach for the interrogation of genes located in ancestry-enriched regions that yielded the examples described above, functional enrichment analysis was used to evaluate whether specific pathways are over-represented among ancestry-enriched genes. To do this, gene set enrichment (GSEA)³⁵ pathway analyses were conducted separately for genomic loci that show African, Asian (Native American) and European enriched ancestry contributions. A number of pathways show up as over-represented when compared against ancestry-enriched genes (Fig. 6A), and similar pathways are found among the different ancestry components despite the fact that their gene sets are mutually exclusive by operational definition. In particular, immune-related pathways consistently appear as significantly over-represented for all three of the ancestry components.

The significant over-representation of ancestry-enriched genes among components of the immune system includes genes that map to pathways involved in both the innate and adaptive immune response (Fig. 6B). The related Toll-like receptor and interferon signaling pathways of the innate immune response contain numerous ancestry-enriched genes, including cytoplasmic members of the NF- κ B and JAK-STAT

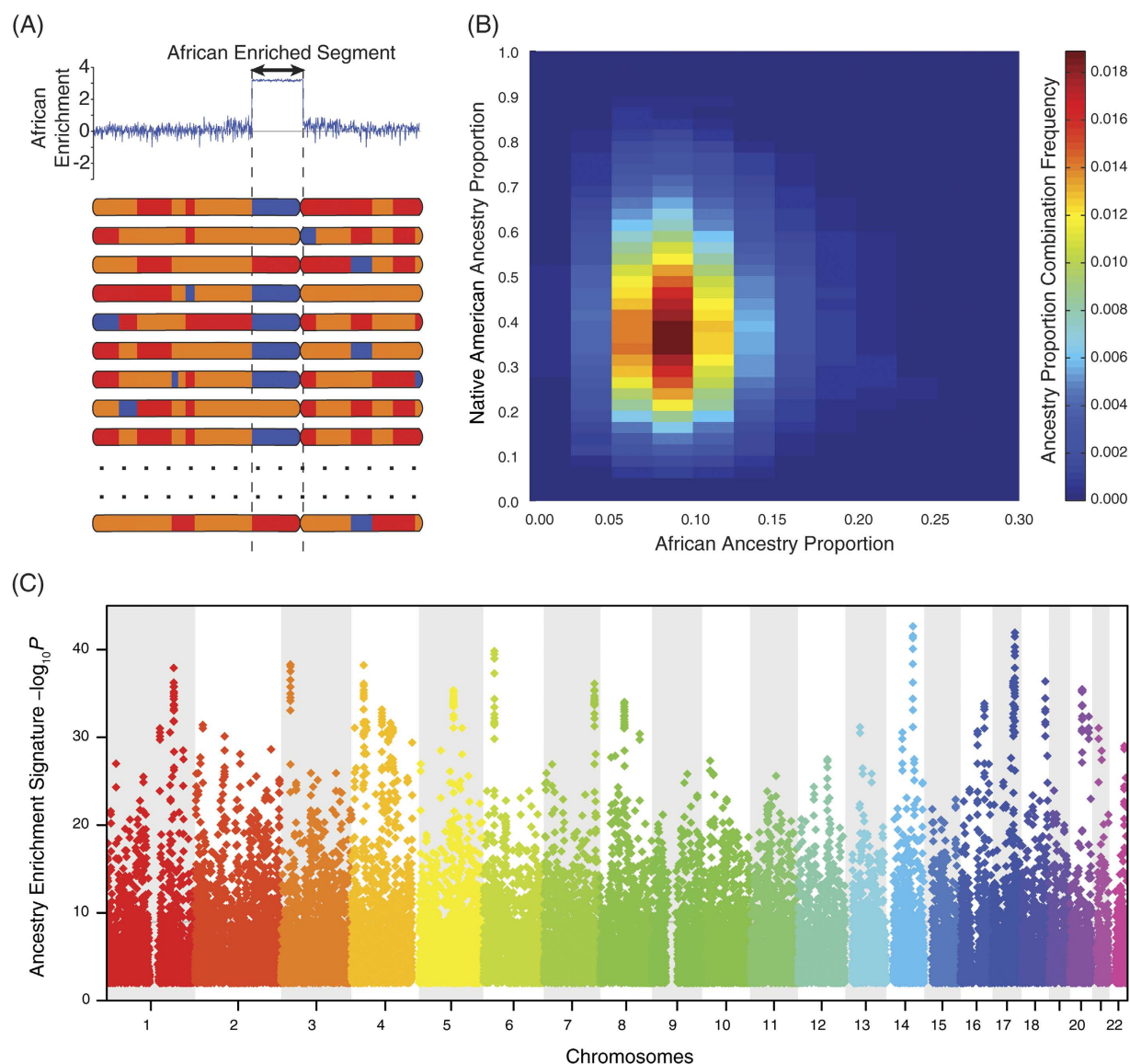


Figure 4. Population-wide ancestry enrichment for Colombian genomes. (A) Schematic of the ancestry enrichment technique. Ancestry-enriched regions are identified as genomic loci with anomalously high ancestry contributions, for the entire population, from one of the three continental components. (B) Heatmap showing the population frequencies of three-way ancestry proportion combinations for Colombian genomes. Each block corresponds to a specific three-way ancestry proportion combination, and the frequency of that combination in the population is color-coded as shown in the scale adjacent to the heatmap. Note that since there are three continental ancestry components, the value of the third ancestry component (European) is dependent on the first two and thus not shown. (C) Manhattan plot showing the genomic regions identified as ancestry-enriched via the trinomial probability (y-axis) of observing a particular three-way ancestry combination in the population (see Materials and Methods).

signaling cascades. There are also ancestry-enriched genes that encode downstream members of these pathways, including transcription factors (NF- κ B), inflammatory cytokines (IL1B) and chemokines (CXCL9, 10 and 11), which together help direct antimicrobial responses via mechanisms such as host cell apoptosis and T-cell chemotaxis. Ancestry-enriched genes of the adaptive immune response encode members of the B-cell and T-cell receptor signaling pathways, including antigen receptors (CD79B) and cytoplasmic signaling molecules (BLNK and RAC1) along with the NF- κ B transcription factor. These proteins help facilitate the proliferation and differentiation of B-cells and T-cells in response to specific immune challenges. Interestingly, these pathways contain a mix of ancestry-enriched genes from different ancestral components. This indicates that individuals from the admixed Colombian population have assembled immune system pathways that are made up of combinations of ancestry-specific alleles that

Gene	Function	Ancestry	Associated trait	Evidence	PMID	Ancestry enrichment (-log ₁₀ P)
ADCY3	Adenylate cyclase 3	Asian	Body mass index	Positive Selection	22344219	18.6
ATM	Ataxia telangiectasia mutated	Asian	Cell cycle	Association Study	24390342, 21983787	15.7
BCHE	Butyrylcholinesterase	African	Cardiovascular disease risk in European	Association Study	23419831, 21943158, 9780523	23.7
CASP8	Caspase 8, apoptosis-related cysteine peptidase	Asian	Reduces breast cancer risk	Positive Selection	17293864	9.6
CD226	CD226 molecule	Asian	Adaptive immune system, cell adhesion	Positive Selection	24390342, 23128233, 21829393, 17554260	34.5
HLA-B	Major histocompatibility complex, (MHC) class I, B	African	Immune response	Positive Selection	23731540, 16998491	31.6
MANBA	Mannosidase	Asian	Decreases colorectal cancer risk & associated with mannosidosis	Association Study	17899454, 21833088	29.7
MAPK10	Mitogen-activated protein kinase 10	Asian	Immune system	Association Study	25634076	25.4
MICA	MHC class I polypeptide-related sequence A	African	Immune response	Positive Selection	23731540	32.6
NANOS3	Nanos homolog 3	Asian	Germ cell development	Association Study	21421998	23.9
NFKB1	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	Asian	Reduces breast cancer risk	Positive Selection	22562547	30.8
SF3B4	Splicing Factor 3b, Subunit 4	Asian	Height	Association Study	20881960, 18391951	30.6
SLC24A5	Solute carrier family 24, member 5	European	Decreases melanin pigmentation in skin	Positive Selection	17182896	16.1
SLC44A2	Solute carrier family 44, member 2	European	Metabolism	Association Study	22040064	20.1
ULBP1	UL16 binding protein 1	European	Cytosolic DNA sensing pathway	Association Study	20923822	12.3
USP32	Ubiquitin Specific Protease 32	Asian	Overexpressed in breast cancer	Positive Selection	19307593	36.4

Table 2. Genes in ancestry-enriched regions and their associated traits.

have never been seen in the same genetic background. This may have provided a mechanism to confront the novel combinations of microbial pathogens found in the New World. It should be noted that gene function tends to be spatially correlated, and this is particularly true for the kinds of immune genes discussed here. As this could lead to a violation of the assumptions of independence that underlie GSEA pathway analysis, the exact *P*-values reported here should be taken with some caution.

Discussion

Genetic ancestry in Colombia. Results from our analysis of 60 Colombian genome sequences from Medellin point to high levels of three-way continental admixture, with contributions from African, Native American and European ancestral source populations, consistent with previous genetic studies^{19–23,25} and with overall demographic trends in Colombia (Table 1). On average, admixed genomes from Medellin show predominantly European ancestry; the average genome sequence shows 74.6% European, 18.1% Asian (Native American) and 7.3% African ancestry (Fig. 1D). However, individuals from Medellin vary widely with respect to their ancestry proportions from these three ancestral source populations. There are a number of individuals with >95% European ancestry on one end of the spectrum and people with far more even three-way contributions from the ancestral populations on the other end (Fig. 1C).

It should be noted that different regions of Colombia show very distinct demographic patterns. For example, whereas people from Medellin have primarily European ancestry, the Atlantic and Pacific coastal regions are home to much larger populations of Afro-Colombians. Thus, the admixture patterns reported here for Medellin can not be taken to represent country-wide patterns of Colombian genetic ancestry. Indeed, studies on the genetic ancestry of Colombians sampled from different regions of the country often yield very different results^{19–23,25}. A paper that was published while our own work was in preparation may represent the most comprehensive survey of Colombian genetic ancestry to date²⁴. In that study, three-way admixture patterns were inferred for 1,659 individuals based on 30 ancestry informative markers. The results of the study underscore the genetic diversity of the Colombian population. The Colombian samples showed the highest levels of average three-way admixture contributions from ancestral populations (60% European, 29% Native American and 11% African) among the five Latin American countries surveyed as well as the greatest extent of geographical variation in genetic ancestry.

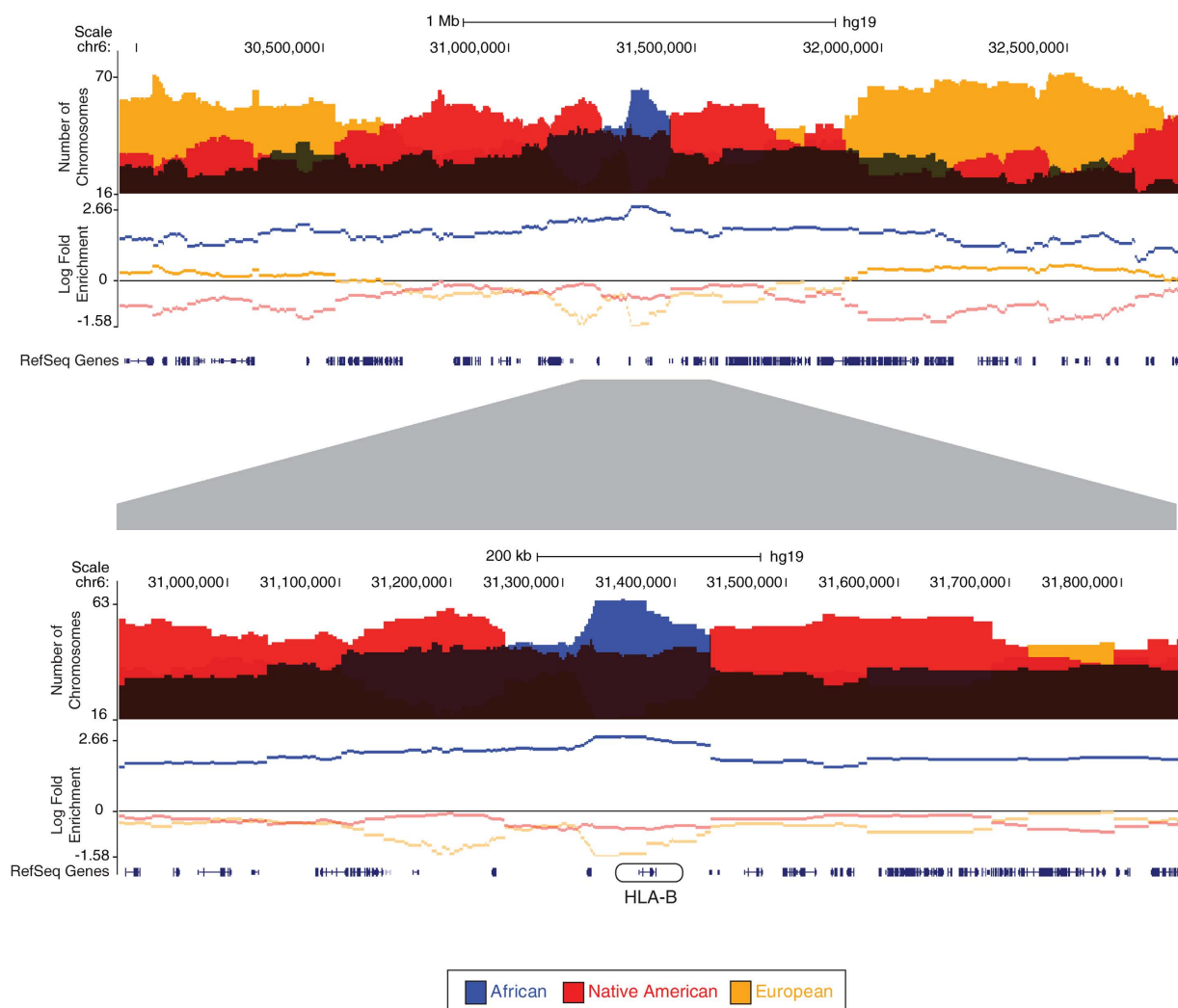


Figure 5. Example of an African-enriched region found in the MHC locus. Population-wide chromosome counts from the three ancestry-components are shown above the genomic axis and log-fold enrichment values (observed counts/genomic average counts) for the ancestry components are shown below the axis. The upper panel shows a ~2.5 Mb region of the MHC locus on chromosome 6, and the lower panel shows a zoomed in view centered on the African ancestry-enriched gene *HLA-B*.

Ancestry and identity in a Colombian population. Interestingly, the genetic ancestry and admixture results obtained here for the population of Medellin can be considered to be at odds with the demographic data for the city. Medellin is considered to have a population that is almost entirely descended from Europeans. In the 2005 census, 93.4% of the population of Medellin was classified as Euro-descendent, whereas 6.5% of individuals identified as Afro-Colombian and only 0.1% identified as Native American³⁶. These demographic data are based on self-identification and reflect the ethnic groups that individuals consider themselves to be members of. Thus, it would appear that the vast majority Colombians from Medellin identify as white despite the presence of a substantial fraction of individuals with appreciable levels of Native American and African ancestry. For example, 46 (~77%) of the Colombian individuals studied here have >2% African ancestry. As 93.4% of the population of Medellin self-identifies as white, it is likely that the majority of these individuals (~43) would self-identify as white. If we assume that these 43 individuals occupy the lower end of the distribution of African ancestry, then 72% of self-identified white Colombians from Medellin have >2% African ancestry. By way of comparison, a recent large-scale analysis of genetic ancestry among different ethnic groups in the United States showed that only 1.4% of self-identified European Americans have at least 2% African ancestry³⁷. It should also be noted that the Colombian individuals studied here have substantially higher levels of Native American compared with African ancestry. However, it is not possible to perform the same kind of comparison between ethnic self-identification and genetic ancestry with respect to Native American ancestry given the city-by-city ethnic category data provided in the Colombian census³⁶. This is because

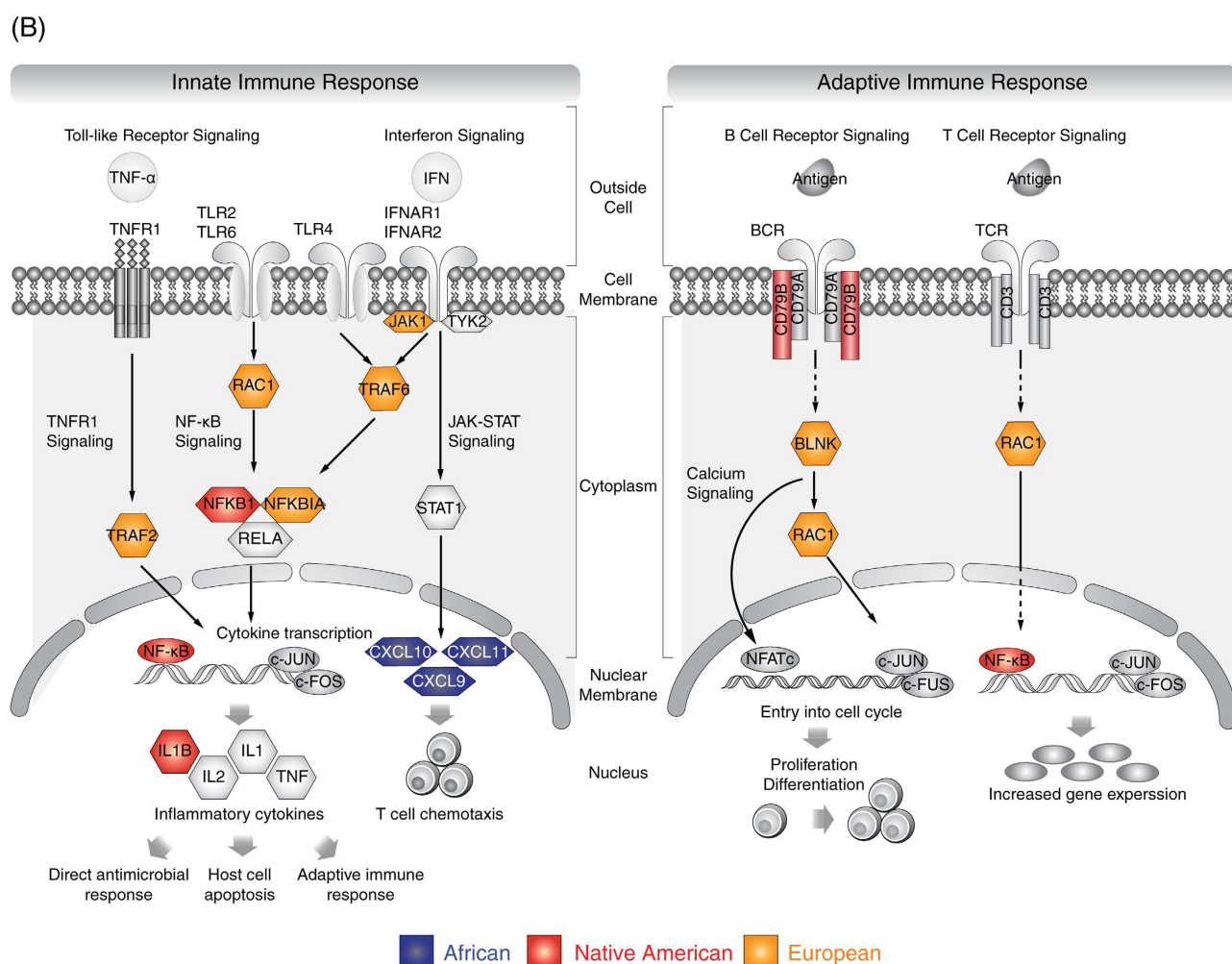
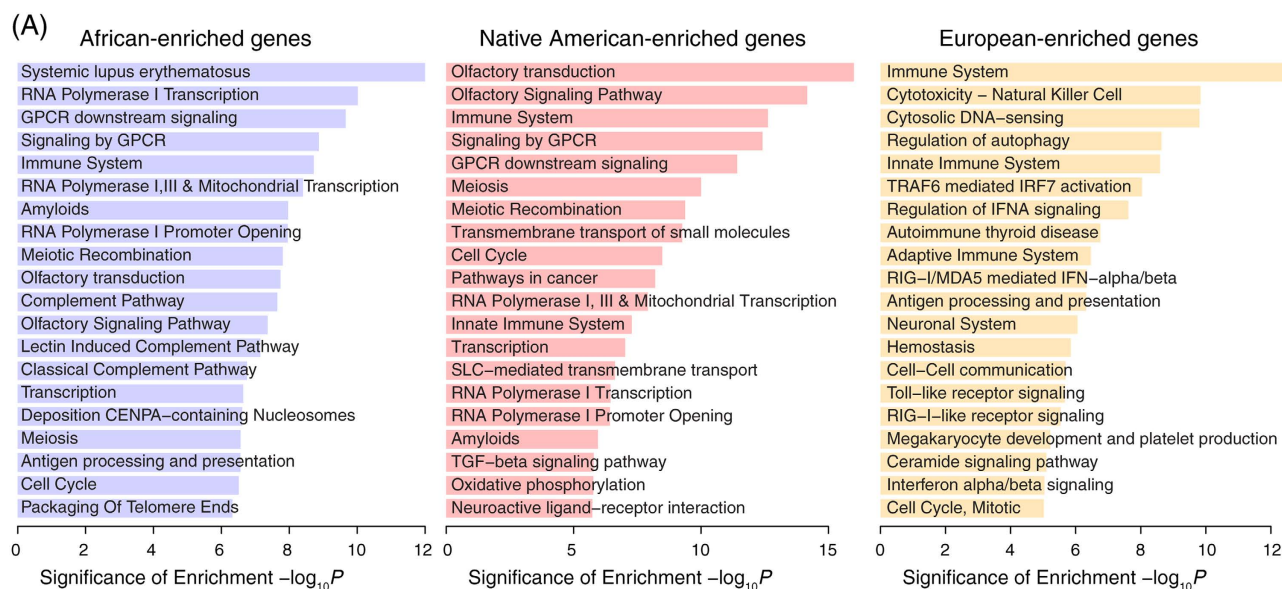


Figure 6. Functional enrichment analysis for ancestry-enriched regions. (A) Pathways identified as over-represented among ancestry-enriched regions via gene set enrichment analysis. (B) Schematic of four such pathways that are involved in innate and adaptive immune response.

the census does not distinguish between European descendants who self-identify as white versus individuals with a combination of European and Native American ancestry who identify as mestizo at the level of individual cities.

The distinction between genetic ancestry and ethnic self-identification in Medellín may be related to two important cultural concepts rooted in many Latin American societies: *Mestizaje* and *Blanqueamiento*³⁸. *Mestizaje* refers to the intentional mixing of different ethnic groups; it is considered to be a critical part of nation-building and cultural identity, for Colombia in particular and across Latin America^{39,40}. *Blanqueamiento* refers to the ideology of racial improvement via the “whitening” of the population. While *blanqueamiento* may have a biological dimension, with respect to the desire to produce whiter offspring, it more often manifests as a social construct. It is in this social sense that *blanqueamiento* is reflected in ethnic self-identification. If whiteness is implicitly held up as a social ideal, and a progressive-generational trend that a society should aspire to, people may tend to self-identify as white irrespective of their genetic ancestry⁴¹. The contrast between the genetic ancestry results obtained here and demographic data point to the possibility that the population of Medellín exemplifies such a trend. On the other hand, since the vast majority of individuals studied here (95%) have genomes with majority European ancestry, these individuals may simply choose to identify most closely as white or European descended despite their genetic admixture.

The relationship between genetic ancestry and ethnic self-identification has recently been studied in depth for five Latin American countries²⁴. This work confirmed that there are large variations in genetic ancestry within self-identified ethnic groups across Latin American countries and reveals the large extent to which physical appearance influences ethnic self-identification. Skin pigmentation was shown to have a particularly profound effect on ethnic self-identification, but physical traits and appearance were also shown to be poor indicators of genetic ancestry in the same study. Nevertheless, the relationship between genetic ancestry and self-identification was shown to be quite complex. There are consistent, albeit weak, correlations between genetic ancestry and physical traits for Latin Americans, and the extent to which genetic ancestry is over- or under-estimated for ethnic groups varies according to both the ethnic category and the particular ancestry component (African, European or Native American). Surprisingly, self-perception of European ancestry tends to underestimate the measured European genetic ancestry, whereas self-perception of African ancestry over-estimates the extent of African genetic ancestry. Native American self-perception underestimates the extent of genetic ancestry at lower levels of Native American genetic ancestry comparable to those observed in this study.

Admixture and the conquest of the Americas. Colombian genome sequences also show strikingly asymmetrical patterns of sex-specific admixture with the male lineages dominated by European ancestry and female lineages comprised of more Native American ancestry (Fig. 2). These results are consistent with a number of previous studies that show similar sex-specific admixture patterns for Colombians^{20–23,25}. But we show here for the first time that the pattern of sex-specific ancestry is more pronounced for the Colombian population than for a number of other Latin American countries. This pattern of genetic ancestry reflects the harsh realities of ‘La Conquista’ and the colonial period that followed. Spanish conquistadors were professional warriors who arrived without their families (wives), and subsequent groups of settlers also included relatively small numbers of women²¹. In addition, armed conflicts between conquistadors and indigenous groups resulted in the extermination of a large part of the Native American population, and men were often targeted for elimination⁴². Together, these demographic and historical factors resulted in European males often having children with indigenous women, accounting for the observed sex-specific admixture trend.

Admixture, fitness and selection in Colombia. When regional-specific ancestry contribution patterns are compared across the entire population of Colombian genomes studied here, it becomes very clear that the patterns of admixture are non-random (Fig. 4). There are numerous ancestry-enriched chromosomal segments that have anomalously high (low) levels of ancestry from one of the three ancestral source populations. These results suggest the possibility that loci enriched for distinct ancestry have been differentially retained in the modern admixed Colombian population owing to their utility in the local environment.

We envision that the particular ‘selection’ process that occurred for the admixed Colombian population, based on assortment among pre-existing ancestry-specific alleles, was somewhat distinct from the concept of natural selection as typically formulated. Adaptive, also referred to as positive or directional, natural selection is most often considered to occur via the fixation of novel mutations based on differential reproductive success. This process starts with the introduction of new alleles by mutation at very low population frequencies and thus is relatively slow; it typically takes place on the order of tens- or hundreds-of-thousands of years³. The evolutionary process that yielded ancestry-enriched segments in the admixed Colombian population was instead based on selection among pre-existing population-specific alleles. These population-specific (or enriched) alleles evolved *in situ* in their ancestral regions, based on local selective pressures, over many thousands of years. When the populations that bore these alleles arrived in the New World, the pre-existing adapted alleles were ready to be rapidly re-assorted into novel admixed genomes. In other words, the long, slow process of natural selection had already occurred in the ancestral source populations to generate a standing pool of genetic variation with

a wide variety of adaptive utility. From this existing pool of genetic variation, numerous ancestry-specific segments containing adaptive alleles were readily accessible for re-assortment. In this way, the enrichment of ancestry-specific segments in the Colombian population, based on their utility in the New World environment, could have occurred much more rapidly, *i.e.* within the relatively short time span following Columbus' arrival in the Americas. The ability to rapidly shape admixed genomes over such a time span, with anomalous combinations of ancestry-specific alleles, is supported by the highly asymmetrical patterns of sex-specific admixture observed for Latino populations.

A similar approach to search for evidence of selection based on ancestry-enrichment has been conducted in two recent studies of admixed African American populations^{43,44}. Results from these two studies differ starkly. The first study reported six highly significant ancestry enriched regions genome-wide and took this as evidence of admixture driven natural selection in African Americans⁴³. The more recent study included a substantially larger sample size along with a stringent control for multiple statistical tests and was not able to replicate the previous finding of significantly enriched ancestry regions⁴⁴. One of the potential limitations of the approach that we use here is the far lower sample size compared with previous studies that employed genotyping arrays (SNP chips). It is possible that the ancestry enriched regions reported here will not be replicated when larger sample sizes are used, as was the case for African Americans. However, our approach does have the advantage of increased resolution for local ancestry inference provided by whole genome sequence comparison, which also necessitates relatively low sample sizes compared to genotype studies. We also employ a different statistical approach for the identification of ancestry enriched regions than used in previous studies, which allows us to look for deviations from expected three-way ancestry patterns. This approach is applied using a stringent control for multiple statistical tests. It is also worth noting that the Colombian population analyzed here, and Latino populations in general, show far more admixture than seen for the African American populations analyzed in the aforementioned studies. It is possible that higher levels of admixture provide more opportunities for the differential retention of previously selected ancestral alleles.

Admixture and the infectious disease burden in Colombia. Interrogation of the functional roles of genes located in the anomalously admixed genomic regions detected here suggests that the retention of ancestry-specific alleles in the Colombian population is most closely tied to immune system function (Table 2 and Fig. 6). This process was likely to have been driven by exposure of the admixed population to numerous microbial pathogens, many of which had never before been encountered. These new pathogens could have come from the tropical environment found in the New World, and/or they could have resulted from the mingling of the three distinct continental population groups, each with their own set of endemic pathogens.

Colombia is located in the tropics, traversed by the equator, and has a high burden of infectious disease caused by a variety of pathogenic agents. Compared with other countries in the region and around the world, Colombia has a very high level of pathogen richness as measured by the number of known pathogens in the country ($n = 244$)^{17,45}. The infectious disease burden in Colombia includes waterborne diseases, such as cholera, and vectorborne diseases, including yellow fever, dengue fever and malaria¹⁸. Host genetic factors play an important role in shaping individuals' susceptibility and resistance to these pathogenic agents⁴⁶. Natural selection has been shown to increase the frequency of alleles that provide increased resistance to cholera, dengue fever and malaria. In addition, many of these alleles are population-specific (or enriched) having evolved within populations from the geographic regions where the diseases originated.

Susceptibility to cholera, which is endemic to South East Asia, is known to be strongly influenced by host genetic factors^{47,48}. These include genes that encode members of the NF- κ B pathway that operates as part of the innate immune system^{49,50}. Results from our analysis show this pathway to be enriched for genes with anomalous ancestry patterns including genes that fall within Asian (Native American) enriched segments (Fig. 6).

There have also been numerous studies that have identified host genetic factors that appear to mitigate susceptibility to dengue fever⁵¹. These include large scale genome-wide association studies (GWAS)^{52–54} along with smaller scale candidate gene studies^{55–57}. A recent GWAS study found the most significant dengue disease association SNPs at the *MICB* locus along with additional significant SNPs at the adjacent *HLA-B* and *HLA-C* loci⁵². All of these genes are members of the MHC locus that encodes numerous cell surface receptor proteins that mediate immunity via the presentation of antigen peptide sequences. All three genes are found within African ancestry-enriched segments (Table 2 and Supplementary Table S1). A number of other candidate gene studies have also uncovered associations between specific *HLA* gene alleles and susceptibility to dengue in Asian populations^{55–57}.

As of this time, there are no known host resistance alleles associated with susceptibility to yellow fever. However, it has long been thought that populations with different genetic ancestries had differing degrees of susceptibility to yellow fever, suggesting a genetic component to susceptibility to this pathogen. In particular, populations with African ancestry were thought to have reduced susceptibility to yellow fever^{58–60}, and this notion has been related to an increased economic incentive for the use of African slave labor in the New World⁶¹. However, this idea has been criticized based on a lack of historical consensus and the absence of direct evidence for a genetic ancestral component to resistance as is seen for malaria⁶². Nevertheless, a recent analysis of mortality from yellow fever in 19th century United States provides

compelling statistical evidence for an association between genetic ancestry and susceptibility to yellow fever⁶³. Individuals with African ancestry were found to have indistinguishable incidence rates of yellow fever compared with Europeans but had significantly lower rates of mortality. The authors went on to rule out a number of environmental (e.g. social and/or economic) factors that could explain this difference and concluded that their findings support the existence of host genetic factors that mitigate the severity of yellow fever infections among individuals with African ancestry.

Malaria poses a particularly grave and consistent threat to public health in Colombia with both endemic and epidemic transmission regimes⁶⁴. While Medellín does not have a high incidence of malaria owing to its altitude, malaria is endemic to nearby regions at lower altitudes, particularly in jungle areas along the Atlantic and Pacific coasts⁶⁵. As the population of Medellín is cosmopolitan³⁶, having consistently received immigrants from surrounding regions over the centuries, malaria may have exerted a selective pressure on the population. Numerous alleles at different genetic loci have evolved under the influence of selective pressure based on resistance to malaria⁴⁶. This process has occurred independently in different populations around the world giving rise to a number of ancestry-specific resistance alleles. For example, there are distinct African- and Asian-specific alleles of the Hemoglobin Beta Chain encoding gene (*HBB*) that evolved independently and provide increased resistance to malaria^{66–68}.

There is a clear connection between the infectious disease burden in Colombia, for a number of the most common pathogen agents found in the country as described above, and host genetic factors that mediate susceptibility and resistance. Many of these host genetic factors are likely to be population-specific (or enriched) alleles that initially evolved within ancestral source populations. The results obtained from our ancestry-enrichment analysis indicate that the modern admixed Colombian population may have had the opportunity to draw from a variety of these ancestry-specific resistance alleles to assemble a diverse repertoire of immune system related genes capable of countering threats from the wide variety of pathogens found in the New World environment. In addition to being of interest from an evolutionary perspective, these results may have implications for better understanding the genetic determinants of health in highly admixed Latin American populations.

Materials and Methods

Genome sequence data. A total of 581 whole genome sequences or genotypes, taken from a variety of sources, were analyzed here (see Supplementary Fig. S1). Whole genome sequences for 60 admixed Colombians from Medellín were taken from the 1000 Genomes Project^{26,27}. Whole genome sequences ($n = 264$) and genotypes characterized using SNP microarrays ($n = 257$) from a number of sources^{25–27,69} were analyzed for putative ancestral populations in Africa, the Americas, Asia and Europe. Genotypes (i.e. SNP calls) characterized using complete genome sequences and microarrays from these different studies (and formats) were merged using the program PLINK v1.90⁷⁰, along with custom scripts, in order to create a single set of merged SNPs across all studies. PLINK was then used to perform linkage disequilibrium pruning on the merged SNP data to produce a reduced set of unlinked SNPs [options: —indep-pairwise 100 25 0.05 —mind 1 —geno 0.01 —hwe .001]. These processes were done separately for genome sequence and genotype data together and for genome sequence data alone.

Ancestry and admixture analysis. Allele sharing distances between pairs of genomes were computed as the fraction of differences between SNP calls. Principal components analysis of the resulting pairwise allele sharing distance matrix was performed using the prcomp program from the R package v3.1.2⁷¹ [options: scale=TRUE] to relate admixed and ancestral genomes. The program ADMIXTURE v1.23⁷² was used to estimate the admixture fractions of three putative ancestral populations – African, Asian (Native American) and European – among Colombian genome sequences. ADMIXTURE was run with default settings and $k = 3$ ancestral populations. The program SupportMix (Ver Jul 18 2012)²⁹ was used to characterize the regional (locus-specific) three-way ancestry admixture fractions in the Colombian genomes using default settings. Locus-specific ancestry admixture analysis was done using whole genome sequences from the 1000 Genomes Project (Supplementary Fig. S1) to afford increased resolution.

Sex-specific admixture contributions. Normalized ratios of the difference between admixture fractions for the X chromosome versus the autosomes were used to infer sex-specific admixture contributions for each of the three ancestral source populations. The admixture difference ratio ($\Delta Admix$) for each ancestry was calculated as:

$$\overline{\Delta Admix} = F_{anc,total} \times (F_{anc,X} - F_{anc,auto}) / (F_{anc,X} + F_{anc,auto}) \quad (1)$$

where $F_{anc,total}$ is the genome-wide admixture proportion for a given ancestry, $F_{anc,X}$ is the X chromosome admixture proportion for a given ancestry and $F_{anc,auto}$ is the autosomal admixture proportion for a given ancestry. Positive values of $\Delta Admix$ are indicative of an excess of female-specific (maternal) admixture contributions, whereas negative values indicate an excess of male-specific (paternal) admixture. The normalized X chromosome versus autosome ratio values were computed for Colombian genomes along with previously reported genome data from four other Latin American countries²⁵. A bootstrapping procedure

(1000 replicates) was used to create pseudo-replicate data sets to compute 95% confidence intervals for the observed ΔAdmix values.

Ancestry enrichment analysis. Chromosomal regions with anomalous patterns of ancestry, when considered as an ensemble across the entire population of 60 Colombian genomes (120 chromosomes), were identified via ancestry-enrichment probability values. These values were computed as the trinomial probability of observing a particular numerical combination of African (*YRI*), Asian (*CHB*) and European (*CEU*) chromosomal segments, for the entire population at a specific genomic locus, given the population-average levels of African, Asian and European ancestry. This probability is expressed as:

$$P(YRI = x_1 \text{ and } CHB = x_2 \text{ and } CEU = x_3) = \frac{120!}{x_1! \times x_2! \times x_3!} p_{YRI}^{x_1} \times p_{CHB}^{x_2} \times p_{CEU}^{x_3} \quad (2)$$

where x_1 , x_2 , x_3 are the observed chromosomal segment counts at a particular locus with *YRI*, *CHB* and *CEU* ancestries, and p_{YRI} , p_{CHB} , p_{CEU} are the expected probabilities based on the population-average ancestry values.

Genomic regions with highly statistically significant ($P < 10^{-9}$, FDR q-value 6.5×10^{-9}) anomalous patterns of ancestry identified in this way were interrogated for their potential functional significance using several sources of information. Genes that lie within these regions were compared to sets of genes previously identified as being subject to positive (adaptive) selection in a number of studies^{6–10,46}, and to genes implicated in human health/disease via association studies using literature surveys and the NHGRI GWAS catalog (accessed May 2014)⁷³. The functional affinities of genes encoded in these regions were also evaluated with gene set enrichment analysis using the GSEA software web v3.87³⁵.

References

- Crosby, A. *The Columbian Exchange: Biological Consequences of 1492*. (Greenwood, 1972).
- Mann, C. C. 1493: *Uncovering the new world Columbus created*. (Alfred a Knopf, 2013).
- Garrigan, D. & Hammer, M. F. Reconstructing human origins in the genomic era. *Nature reviews. Genetics* **7**, 669–680, doi: 10.1038/nrg1941 (2006).
- Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104, doi: 10.1126/science.1153717 (2008).
- Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385, doi: 10.1126/science.1078311 (2002).
- Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome research* **19**, 826–837, doi: 10.1101/gr.087577.108 (2009).
- Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918, doi: 10.1038/nature06250 (2007).
- Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620, doi: 10.1126/science.1124309 (2006).
- Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713, doi: 10.1016/j.cell.2013.01.035 (2013).
- Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS genetics* **3**, e90, doi: 10.1371/journal.pgen.0030090 (2007).
- Soejima, M., Tachida, H., Ishida, T., Sano, A. & Koda, Y. Evidence for recent positive selection at the human AIM1 locus in a European population. *Molecular biology and evolution* **23**, 179–188, doi: 10.1093/molbev/msj018 (2006).
- Norton, H. L. *et al.* Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular biology and evolution* **24**, 710–722, doi: 10.1093/molbev/msl203 (2007).
- Myles, S., Somel, M., Tang, K., Kelso, J. & Stoneking, M. Identifying genes underlying skin pigmentation differences among human populations. *Human genetics* **120**, 613–621, doi: 10.1007/s00439-006-0256-4 (2007).
- McEvoy, B., Belez, S. & Shriver, M. D. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Human molecular genetics* **15 Spec No 2**, R176–181, doi: 10.1093/hmg/ddl217 (2006).
- Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A. & Kayser, M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human genetics* **71**, 354–369, doi: 10.1111/j.1469-1809.2006.00341.x (2007).
- Izagirre, N., Garcia, I., Junquera, C., de la Rua, C. & Alonso, S. A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular biology and evolution* **23**, 1697–1706, doi: 10.1093/molbev/msl030 (2006).
- Sanchez-Mazas, A., Lemaitre, J. F. & Currat, M. Distinct evolutionary strategies of human leucocyte antigen loci in pathogen-rich environments. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **367**, 830–839, doi: 10.1098/rstb.2011.0312 (2012).
- CIA. CIA World Factbook: Colombia. (2014). <<https://www.cia.gov/library/publications/the-world-factbook/geos/co.html>>. Date of access: 22/04/2015.
- Wang, S. *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS genetics* **4**, e1000037, doi: 10.1371/journal.pgen.1000037 (2008).
- Cordoba, L. *et al.* Composicion genetica de una poblacion del suroccidente de Colombia. *Revista Colombiana de Antropologia* **48**, 21–48 (2012).
- Carvajal-Carmona, L. G. *et al.* Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *American journal of human genetics* **67**, 1287–1295, doi: 10.1016/S0002-9297(07)62956-5 (2000).
- Carvajal-Carmona, L. G. *et al.* Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Human genetics* **112**, 534–541, doi: 10.1007/s00439-002-0899-8 (2003).
- Bedoya, G. *et al.* Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 7234–7239, doi: 10.1073/pnas.0508716103 (2006).
- Ruiz-Linares, A. *et al.* Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS genetics* **10**, e1004572, doi: 10.1371/journal.pgen.1004572 (2014).

25. Bryc, K. *et al.* Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America* **107** Suppl 2, 8954–8961, doi: 10.1073/pnas.0914618107 (2010).
26. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, doi: 10.1038/nature11632 (2012).
27. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073, doi: 10.1038/nature09534 (2010).
28. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370–374, doi: 10.1038/nature11258 (2012).
29. Omberg, L. *et al.* Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC genetics* **13**, 49, doi: 10.1186/1471-2156-13-49 (2012).
30. Garamszegi, L. Z. Global distribution of malaria-resistant MHC-HLA alleles: the number and frequencies of alleles and malaria risk. *Malaria journal* **13**, 349, doi: 10.1186/1475-2875-13-349 (2014).
31. Liu, X. *et al.* Differential positive selection of malaria resistance genes in three indigenous populations of Peninsular Malaysia. *Human genetics*, doi: 10.1007/s00439-014-1525-2 (2015).
32. Aoki, K. Sexual selection as a cause of human skin colour variation: Darwin's hypothesis revisited. *Annals of human biology* **29**, 589–608, doi: 10.1080/0301446021000019144 (2002).
33. Madrigal, L. & Kelly, W. Human skin-color sexual dimorphism: a test of the sexual selection hypothesis. *American journal of physical anthropology* **132**, 470–482, doi: 10.1002/ajpa.20453 (2007).
34. Kamberov, Y. G. *et al.* Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702, doi: 10.1016/j.cell.2013.01.016 (2013).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550, doi: 10.1073/pnas.0506580102 (2005).
36. DANE. *Censo General 2005: Perfil Medellin Antioquia*, <http://www.dane.gov.co/files/censo2005/PERFIL_PDF_CG2005/050017T7000.PDF> (2005). Date of access: 22/04/2015.
37. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *American journal of human genetics* **96**, 37–53, doi: 10.1016/j.ajhg.2014.11.010 (2015).
38. Wade, P. Blackness and race mixture: the dynamics of racial identity in Colombia. (JHU Press, 1995).
39. Chavez, M. & Zambrano, M. in *European Review of Latin American and Caribbean Studies* Vol. **80** 5–23 (2006).
40. Andrews, G. R. Afro-Latin America 1800–2000. (Oxford University Press, 2004).
41. Telles, E. & Flores, R. Not just color: Whiteness, nation, and status in Latin America. *Hispanic American Historical Review* **93**, 411–449 (2013).
42. Markham, C. The Conquest of New Granada. (EP Dutton and Company, 1912).
43. Jin, W. *et al.* Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome research* **22**, 519–527, doi: 10.1101/gr.124784.111 (2012).
44. Bhatia, G. *et al.* Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *American journal of human genetics* **95**, 437–444, doi: 10.1016/j.ajhg.2014.08.011 (2014).
45. Informatics, G. *The world's premier global infectious diseases database*. <<http://www.gideononline.com/>> (2015). Date of access: 22/04/2015.
46. Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nature reviews. Genetics* **15**, 379–393, doi: 10.1038/nrg3734 (2014).
47. Harris, J. B. *et al.* Susceptibility to *Vibrio cholerae* infection in a cohort of household contacts of patients with cholera in Bangladesh. *PLoS neglected tropical diseases* **2**, e221, doi: 10.1371/journal.pntd.0000221 (2008).
48. Barua, D. & Paguio, A. S. ABO blood groups and cholera. *Annals of human biology* **4**, 489–492 (1977).
49. Lee, P. H., O'Dushlaine, C., Thomas, B. & Purcell, S. M. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* **28**, 1797–1799, doi: 10.1093/bioinformatics/bts191 (2012).
50. Karlsson, E. K. *et al.* Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. *Science translational medicine* **5**, 192ra186, doi: 10.1126/scitranslmed.3006338 (2013).
51. Stephens, H. A. HLA and other gene associations with dengue disease severity. *Current topics in microbiology and immunology* **338**, 99–114, doi: 10.1007/978-3-642-02215-9_8 (2010).
52. Khor, C. C. *et al.* Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nature genetics* **43**, 1139–1141, doi: 10.1038/ng.960 (2011).
53. Whitehorn, J. *et al.* Genetic variants of MICB and PLCE1 and associations with non-severe dengue. *PloS one* **8**, e59067, doi: 10.1371/journal.pone.0059067 (2013).
54. Devignot, S. *et al.* Genome-wide expression profiling deciphers host responses altered during dengue shock syndrome and reveals the role of innate immunity in severe dengue. *PloS one* **5**, e11671, doi: 10.1371/journal.pone.0011671 (2010).
55. Stephens, H. A. *et al.* HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. *Tissue antigens* **60**, 309–318 (2002).
56. Nguyen, T. P. *et al.* Protective and enhancing HLA alleles, HLA-DRB1*0901 and HLA-A*24, for severe forms of dengue virus infection, dengue hemorrhagic fever and dengue shock syndrome. *PLoS neglected tropical diseases* **2**, e304, doi: 10.1371/journal.pntd.0000304 (2008).
57. Appanna, R., Ponnampalavanar, S., Lum Chai See, L. & Sekaran, S. D. Susceptible and protective HLA class I alleles against dengue fever and dengue hemorrhagic fever patients in a Malaysian population. *PloS one* **5**, doi: 10.1371/journal.pone.0013029 (2010).
58. Kiple, K. F. The Caribbean slave: a biological history. (Cambridge University Press Cambridge, 1984).
59. Kiple, K. F. & Kiple, V. H. Black yellow fever immunities, innate and acquired, as revealed in the American South. *Social Science History* **1**, 419–436 (1977).
60. McNeill, J. R. Ecology, epidemics and empires: environmental change and the geopolitics of tropical America, 1600–1825. *Environment and History* **5**, 175–184 (1999).
61. McNeill, J. R. Mosquito Empires: ecology and war in the Greater Caribbean. 1620–1914. (Cambridge University Press, 2010).
62. Espinosa, M. The question of racial immunity to Yellow Fever in history and historiography. *Social Science History*. in press.
63. Blake, L. E. & Garcia-Blanco, M. A. Human genetic variation and yellow fever mortality during 19th century U.S. epidemics. *mBio* **5**, e01253–01214, doi: 10.1128/mBio.01253-14 (2014).
64. Ramirez, A. P., Buitrago, J. I., Gonzalez, J. P., Morales, A. H. & Carrasquilla, G. Frequency and tendency of malaria in Colombia, 1990 to 2011: a descriptive study. *Malaria journal* **13**, 202, doi: 10.1186/1475-2875-13-202 (2014).
65. Valero, M. V., Vergara, J. & Herrera, S. Malaria Atlas Project. (2015). <<http://www.map.ox.ac.uk/explore/countries/COL/>>. Date of access: 22/04/2015.

66. Ohashi, J. *et al.* Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *American journal of human genetics* **74**, 1198–1208, doi: 10.1086/421330 (2004).
67. Modiano, D. *et al.* Haemoglobin C protects against clinical *Plasmodium falciparum* malaria. *Nature* **414**, 305–308, doi: 10.1038/35104556 (2001).
68. Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96**, 2358–2363 (2000).
69. Watkins, W. S. *et al.* Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. *BMC genetics* **13**, 39, doi: 10.1186/1471-2156-13-39 (2012).
70. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575, doi: 10.1086/519795 (2007).
71. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* **5**, 299–314 (1996).
72. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655–1664, doi: 10.1101/gr.094052.109 (2009).
73. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001–1006, doi: 10.1093/nar/gkt1229 (2014).

Author Contributions

L.R. and A.B.C. conducted all of the comparative genome sequence analyses including the ancestry and admixture analysis. C.H.W. developed and applied the ancestry enrichment analysis method. L.W. conducted the gene set and pathway enrichment analyses. A.V.A. and I.K.J. conceived of, designed and supervised all aspects of the study. I.K.J., L.R. and A.V.A. wrote and revised the main manuscript text. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Rishishwar, L. *et al.* Ancestry, admixture and fitness in Colombian genomes. *Sci. Rep.* **5**, 12376; doi: 10.1038/srep12376 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>



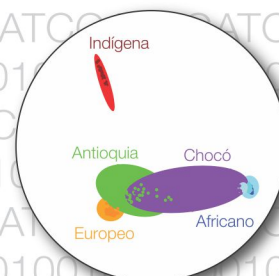
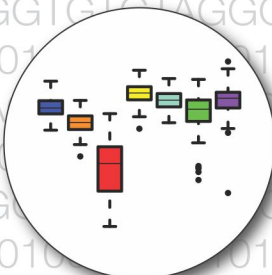
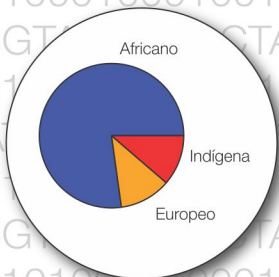
Revista Biodiversidad Neotropical

ISSN 2027-8918

Journal Neotropical Biodiversity / Rev. Biodivers. Neotrop.

Revista Internacional de la Biota Neotropical /International Journal of the Neotropical Biota

Evidencias de la diversidad genética de la población afrodescendiente en el Chocó, Colombia.



CHOCOGEN
Patrimonio Genético del
Pacífico Colombiano



Volúmen 6



Universidad Tecnológica del Chocó
Diego Luis Córdoba

No. 1 Enero - Junio de 2016

Volumen 6 No. 1 January-June de 2016

Chocó, Colombia: a hotspot of human biodiversity

El Chocó, Colombia: un *hotspot* de la biodiversidad humana

Miguel A. Medina-Rivas^{1,2}, Emily T. Norris³, Lavanya Rishishwar^{3,4}, Andrew B. Conley³,
Camila Medrano-Trochez⁵, Augusto Valderrama-Aguirre^{2,5}, Fredrik O. Vannberg⁴,
Leonardo Mariño-Ramírez^{2,6,7}, I. King Jordan^{2,4,7}

Abstract

Objective: Chocó is a state located on the Pacific coast of Colombia that has a majority Afro-Colombian population. The objective of this study was to characterize the genetic ancestry, admixture and diversity of the population of Chocó, Colombia. **Methodology:** Genetic variation was characterized for a sample of 101 donors (61 female and 40 male) from the state of Chocó. Genotypes were determined for each individual via the characterization of 610,545 single nucleotide polymorphisms genome-wide. Haplotypes for the uniparental mitochondrial DNA (female) and Y-DNA (male) chromosomes were also determined. These data were used for comparative analyses with a number of worldwide populations, including putative ancestral populations from Africa, the Americas and Europe, along with several admixed American populations. **Results:** The population of Chocó has predominantly African genetic ancestry (75.8%) with approximately equal parts European (13.4%) and Native American (11.1%) ancestry. Chocó shows relatively high levels of three-way genetic admixture, and far higher levels of Native American ancestry, compared to other New World African populations from the Caribbean and the United States. There is a striking pattern of sex-specific ancestry in Chocó, with Native American admixture along the female lineage and European admixture along the male lineage. The population of Chocó is also characterized by relatively high levels of overall genetic diversity compared to both putative ancestral populations and other admixed American populations. **Conclusion:** These results suggest a unique genetic heritage for the population of Chocó and underscore the profound human genetic diversity that can be found in the region.

Keywords: Admixture, Afro-Colombian, Colombia, Genetic ancestry, Genetic diversity, Human genome.

Resumen

Objetivo: El Chocó es un departamento situado en la costa pacífica colombiana cuya población es mayoritariamente afrocolombiana. El objetivo de este estudio fue caracterizar la ancestralidad, el mestizaje y la diversidad genética de la población del Chocó colombiano. **Metodología:** La variación genética se caracterizó en una muestra de 101 donantes (61 mujeres y 40 hombres) del departamento del Chocó. Los genotipos se determinaron para cada individuo a través de la caracterización de 610,545 polimorfismos de nucleótido único distribuidos en todo el genoma. También se determinaron los haplotipos de los ADNs uniparentales: ADN mitocondrial (materno) y cromosoma Y (paterno). Estos datos se utilizaron para análisis comparativos en una serie de poblaciones de todo el mundo, incluyendo poblaciones ancestrales putativas de África, América y Europa, además de varias poblaciones mestizas americanas. **Resultados:** La población del Chocó tiene una ancestralidad genética predominantemente africana (75,8%), con aportes similares de ancestralidad europea (13,4%) y de nativos americanos (11,1%). La población analizada del Chocó tiene niveles relativamente altos de mestizaje triple y niveles mucho más altos de ancestralidad nativa americana, en comparación con otras poblaciones afrodescendientes del nuevo mundo, localizadas

¹ Centro de Investigación en Biotecnología y Recursos Fitogenéticos. Centro de Investigaciones en Biodiversidad y Hábitat, Universidad Tecnológica del Chocó, Quibdó, Chocó, Colombia.

² PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia.

³ IHRC-Georgia Tech Applied Bioinformatics Laboratory, Atlanta, Georgia, USA.

⁴ School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA. e-mail: king.jordan@biology.gatech.edu

⁵ Biomedical Research Institute, Universidad Libre, Cali, Valle del Cauca, Colombia.

⁶ National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland, USA.

⁷ BIOS Centro de Bioinformática y Biología Computacional, Manizales, Caldas, Colombia.

Date received: October 16, 2015

Date approval: January 20, 2016

Associated Editor: Jimenez AM.

en el Caribe y los Estados Unidos. Hay un patrón fuerte de ancestralidad sexo-específica en el Chocó, con un mestizaje predominantemente nativo americano en el linaje femenino y europeo en el linaje masculino. La población del Chocó también se caracteriza por niveles relativamente altos de diversidad genética global en comparación con las poblaciones ancestrales putativas y otras poblaciones mestizas americanas. **Conclusión:** Estos resultados sugieren un patrimonio genético único para la población del Chocó y subrayan la profunda diversidad genética humana que se puede encontrar en la región.

Palabras clave: Afrocolombianos, Ancestralidad genética, Colombia, Diversidad genética, Genoma humano, Mestizaje.

Introduction

Chocó, Colombia. Chocó is a Colombian administrative department (*i.e.*, a state) located along the country's Pacific coast (Figure 1A). Chocó ranges from the Panamanian border in the north to the Cauca Valley in the southwest region of Colombia. The state of Chocó is recognized worldwide as a hotspot of biodiversity (<http://www.eoearth.org/view/article/150631/>). A biodiversity hotspot is defined as a specific geographic region with a large amount of endemic biodiversity that is threatened by human activity (Zachos and Habel 2011). In order for an area to be officially recognized as a biodiversity hotspot, at least 0.5% of its vascular plant species (or 1,500 species) must be characterized as endemics, which are defined as species found uniquely within a proscribed geographic region or habitat type. There are 25 global regions that qualify as biodiversity hotspots

according to this criterion, and together these areas are home to almost 60% of the world's known species of plants, birds, reptiles, amphibians and mammals. The entire area of Chocó is contained within the so-called Chocó-Darién biodiversity hotspot, also known as the Tumbes-Chocó-Magdalena biodiversity hotspot (<http://tmalliance.org/about/where-we-work/biodiversity-hotspot/>).

The Chocó-Darién biodiversity hotspot extends along the Pacific coast from the Panamá Canal region in the north, through the Darién Gap and Chocó wet rain forests in Panamá-Colombia, passing completely through Ecuador before ending in the coastal dry forests of northern Perú. This area encompasses a wide variety of diverse habitats including the wettest rain forests in the world, which are found in Chocó. The Colombian portion of this biodiversity hotspot in Chocó is relatively preserved compared to the Ecuadorean zone where 98% of the native forest has been cleared. The Chocó-Darién biodiversity hotspot supports ~10,000 species of vascular plants along with 600 species of birds, 235 species of mammals, 350 species of amphibians and 210 species of reptiles.

Human biodiversity in Chocó. Although Chocó is widely recognized as a hotspot of biodiversity, one critical aspect of the region's biodiversity remains largely unexplored, the diversity of its human population (Figure 1B). The population of Chocó has a uniquely African genetic heritage with admixture from the Americas and Europe. The vast majority of the population is Afro-Colombian (82.1%) but there are also substantial numbers of Native Americans



Figure 1. Human biodiversity in Chocó. (A) The state of Chocó on the Pacific coast of Colombia. (B) The people of Chocó.

(12.7%) and individuals with primarily European ancestry (5.2%) (Hernández Romero 2005). There are anywhere from 9 to 20 million Afro-descendants in Colombia, making it the country with the third most Afro-descendants in the Americas. Despite the presence of such a large population of Afro-descendants in Colombia, there is a vast under-representation of genetic studies of Afro-Colombians (Rishishwar *et al.* 2015a). Previous Colombian genetic ancestry studies have dealt mainly with Mestizo genomes that have primarily European and secondary Native American ancestry (Carvajal-Carmona *et al.* 2000, Carvajal-Carmona *et al.* 2003, Bedoya *et al.* 2006, Wang *et al.* 2008, Bryc *et al.* 2010, Córdoba *et al.* 2012, Ruiz-Linares *et al.* 2014, Rishishwar *et al.* 2015b). Given the high percentage of Afro-Colombians living in Chocó, genetic studies of this population are ideally suited to uncover the as yet untapped African dimension of Colombian ancestry and human biodiversity.

Colombian individuals have three-way genetic admixture patterns that result from ancestral contributions to the modern population from Africa, the Americas and Europe (Rishishwar *et al.* 2015b). The story of every Colombian's ancestry, along with their specific admixture patterns, is written in the sequence of their genome. Thus, genome sequence analysis can be used to infer genetic ancestry and admixture patterns for individuals and for the population as a whole. For individuals, the total overall proportions of African, Native American and European ancestry can be inferred along with both locus-specific and sex-specific patterns of ancestry and admixture. Genetic ancestry can be explored at both the continental level to uncover the broad regions of origin for Colombians and at the sub-continental level to explore the specific ancestral regions and countries from which individuals' ancestors originate. At the population level, locus-specific admixture patterns can reveal whether natural selection has enriched for specific ancestry along particular genomic segments. Population level inference can also be used to assess whether there are sex-specific differences in ancestry that result from differential ancestry contributions along maternal versus paternal lineages. These kinds of genomic research approaches will be applied in order to elucidate the patterns of genetic ancestry and admixture of the population of Chocó under the auspices of a newly formed collaborative research

project ChocoGen (<http://www.chocogen.com/>).

ChocoGen project. An exploration of human genetic biodiversity in Chocó is being conducted via the collaborative ChocoGen research project in an effort to value, conserve and utilize this precious resource. The ChocoGen research project has two overarching goals:

- 1) to characterize the genetic diversity and ancestry of the population of Chocó, and
- 2) to create a health profile of the region based on the genetic diversity of its people.

Research and development activities in support of both of these goals are being conducted in such a way as to develop the local human capacity in Chocó for research and education in genetic health and medicine. This project is a collaboration between Universidad Tecnológica del Chocó (UTCH) in Colombia, principle investigator Dr. Miguel A. Medina Rivas, and the Georgia Institute of Technology in the USA, principle investigator Dr. I. King Jordan. Bioinformatics analysis and interpretation of human genome sequences from the population of Chocó are being further supported by the National Center for Biotechnology Information (NCBI) in the USA, and the Colombian National Center for Bioinformatics and Computational Biology (BIOS) in Manizales.

Researchers from the ChocoGen project are conducting analysis of genomic sequences sampled from volunteers from the population of Chocó to characterize:

- 1) their genetic ancestry,
- 2) the quantity and nature of genetic admixture between ancestral populations, and
- 3) the possible relationship between genetic ancestry, admixture and determinants of health and disease.

The results of this project will serve as a resource for the development and application of genetic approaches to healthcare in the Pacific region of Colombia and help to position UTCH as a leader in this area of applied research. In this report, we present results of the first round of analyses of the genetic ancestry of 101 individuals sampled from the population of Chocó.

Methodology

Sample donors and genotyping. ChocoGen volunteer DNA sample donors were recruited at UTCH. Donors were selected in an effort to include representative samples of different geographic regions of Chocó (Atrato, Baudó, Atlantic coast, Pacific coast, San Juan), as well as an approximately equal representation of males and females, and donors were asked to self-identify their ethnic origins. Donors contributed DNA using a non-invasive saliva sampling method. All donors signed informed consent documents indicating their understanding of the potential risks of the project along with how their data would be handled and how their identity would be protected. Collection, genotyping and comparative analyses of human DNA samples were conducted with the approval of the ethics committee of UTCH. Donor DNA samples were genotyped using the Illumina HumanOmniExpress-24 single nucleotide polymorphism (SNP) chip.

Comparative genomic data sources. The genotypes of ChocoGen donors were compared to whole genome sequence data from the 1000 Genomes Project (1000G) (Genomes Project *et al.* 2010, Genomes Project *et al.* 2015) and genotype data from the Human Genome Diversity Project (HGDP) (Cann *et al.* 2002, Li *et al.* 2008) (Table 1). Genotypes from the ChocoGen donors, along with genotypes from the 1000G and HGDP projects, were all mapped to the coordinate space of the February 2009 human genome reference sequence version GRCh37/hg19 (Lander *et al.* 2001, Kent *et al.* 2002) for subsequent analysis. The program PLINK (Purcell *et al.* 2007) was used for genotype quality control and to extract autosomal genotyped positions (*i.e.*, single nucleotide polymorphisms or SNPs) common to all three genotype sources to yield a final merged genotype dataset. For quality control, only individual SNP positions with a genotyping rate $\geq 98\%$ were retained for subsequent analysis.

Ancestry and admixture inference. The program PLINK was used to prune the final merged genotype dataset by removing correlated sets of SNPs. Genomic distances were computed as pairwise allele sharing distances between all individual pruned genotypes using PLINK. The resulting pairwise distance matrix was projected onto two-dimensions with principal

component analysis (PCA) using the `prcomp` function from the R package for statistical computing (Team 2008). The program ADMIXTURE (Alexander *et al.* 2009) was run on the genotype dataset to infer individual ancestry components. The resulting data was used with the `nnls` package from R to implement a non-negative least squares method to estimate the fractions of African, Native American and European ancestry for each individual from Chocó. For each individual, the entropy (H) of the admixture was calculated as

$$H = -\sum_{i=1}^3 p_i \times \log(p_i)$$

where p_i is the ancestry fraction for population i . Sex-specific ancestry in the population of Chocó was determined via analysis of uniparental haplotypes: mitochondrial DNA (mtDNA) for the maternal lineage and Y chromosomal DNA (Y-DNA) for the paternal lineage.

The relative genetic diversity levels of the populations analyzed here were measured via the total amount of observed pairwise allele sharing distance and genome-wide heterozygosity. The total amount of observed pairwise allele sharing distance within each population was computed by fitting a minimum spanning ellipse to the individual genotype points of the population projected onto the first two principal components of the PCA analysis using the `ellipsoidhull` function in R. The areas (A) of the population-specific ellipses were computed using the lengths of the major (x) and minor (y) axes scaled to the principal component weights:

$$A = \pi(0.665 * x)(0.060 * y)$$

Heterozygosity was measured as the fraction of all genotype positions that are heterozygous within an individual using the program PLINK. To do this, we analyzed SNPs with minor allele frequency $\geq 25\%$ in order allow for comparison between SNPs called from genome sequences versus SNPs called from genotype arrays, which are biased to high minor allele frequencies and European populations.

Results and discussion

Genetic characterization of the population of Chocó. Volunteer DNA sample donors were solicited on the main campus of UTCH located in the capital

city of Quibdó; 101 volunteers (61 females and 40 males) provided DNA samples for genetic characterization along with answers to a series of questions related to their ethnic self-identity and family history. DNA samples were characterized in order to determine the specific identity of genetic sequence variants at 610,545 loci across the genome. Genetic variants are referred to here as single nucleotide polymorphisms (SNPs), and the specific identity of the DNA sequence residues that correspond to a genome-wide collection of SNPs is referred to as a genotype. The specific identify of the DNA sequence residues for a set of genetically linked SNPs is referred to as a haplotype. For the purposes of this study, donors' genotypes were characterized for the entire set of human autosomes, and haplotypes were determined for uniparental mitochondrial DNA (mtDNA) and Y-DNA chromosomes. Chocó genotypes were compared to genotypes for a variety of human populations (Table 1), in order to make inferences about the genetic ancestry and diversity of the population. Chocó mtDNA and Y-DNA haplotypes were compared to known global distributions for haplotypes of these chromosomes in order to make inferences about female-specific (mtDNA) and male-specific (Y-DNA) genetic ancestry of the population.

Genetic ancestry and admixture of Chocó. The Colombian population has a mixture of genetic ancestry from African, European and Native American pop-

ulations, owing to the historical patterns of conquest and colonization in the New World (Markham 1912, Mann 2013). Thus, Chocó genotypes were compared to genotypes characterized from individuals sampled from representative populations of these regions (Table 1) in order to infer their overall genetic ancestry and admixture. The genetic relationships among individuals from the population of Chocó, along with individuals from the other global populations, are visually represented in Figure 2A. This panel shows a two-dimensional principal component analysis (PCA) projection of the pairwise genetic distances between all of the genotypes analyzed here, where the distance between each dot corresponds to the distance between each individual genotype. The main component of human genetic diversity in this representation is projected along the x-axis (PC1=66.5% of the diversity) and the secondary component is shown on the y-axis (PC2=6%). African, European and Native American populations occupy the three poles of human genetic diversity in this plot, whereas admixed American genomes, including Chocó (CHO) individuals, occupy intermediate positions between these three ancestral, and relatively non-admixed, population groups. The relative positions of the admixed American populations compared to the three ancestral groups gives an indication of their admixture proportions. For example, the Colombian population from Medellín (CLM) shows evidence of more European admixture

Table 1. Human populations analyzed in this study*

Dataset	Color	Short	Full Description	n
<i>Non-admixed Ancestral Populations</i>				
1000G African (n=292)		LWK	Luhya in Webuye, Kenya	99
		MSL	Mende in Sierra Leone	85
		YRI	Yoruba in Ibadan, Nigeria	108
1000G European (n=198)		GBR	British in England and Scotland	91
		IBS	Iberian Population in Spain	107
HGDP Native American (n=70)		KRT	Karitiana in Brazil	24
		PMA	Pima in Mexico	25
		SUR	Surui in Brazil	21
<i>Admixed American Populations</i>				
100G Admixed American (n=251)		ACB	African Caribbean in Barbados	96
		ASW	Americans of African Ancestry in SW USA	61
		CLM	Colombians from Medellin, Colombia	94
Chocó (n=101)		CHO	Colombians from Chocó, Quibdó, Colombia	101

* The sources of the data-1000 Genomes Project (1000G) or Human Genome Diversity Project (HGDP) -are shown along with the population color codes, short abbreviations, full descriptions and numbers (n) of individuals analyzed. Populations are organized according to whether they represent non-admixed ancestral populations from Africa, Europe and the Americas or admixed American populations from the Caribbean, USA and Colombia. Chocó refers to the population sample studied here.

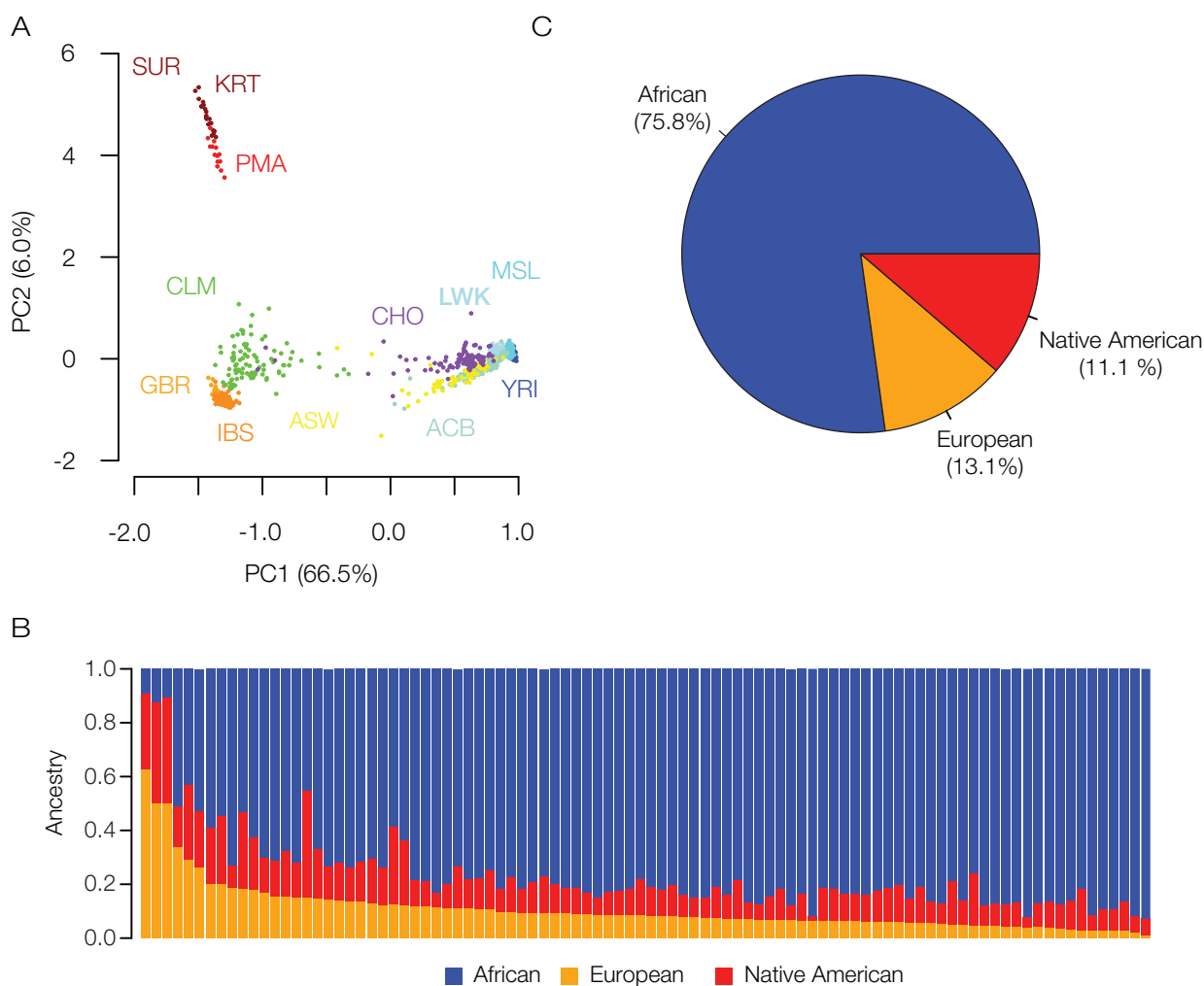


Figure 2. Genetic ancestry and admixture in Chocó. (A) PCA plot showing the genetic relationships among individuals from different human populations compared to the population of Chocó. Each dot corresponds to a single individual, and the distances between dots correspond to the genetic distances between individuals. Populations are colored coded and labeled as shown in Table 1; individuals from Chocó are labeled with CHO and shown in purple. The principal components are labeled (PC1 and PC2) and shown along with the amounts of genetic variation captured by each component. (B) Admixture bar chart showing the percentage of African (blue), European (orange) and Native American (red) ancestry for each individual from population of Chocó. (C) Chocó population-average values for African (blue), European (orange) and Native American (red) ancestry.

compared to the Chocó population, which is located in much closer proximity to the African populations. Two other New World African populations (ACB and ASW) are also located in close proximity to the putative ancestral populations from the African continent, but occupy different positions than the Chocó population.

These same pairwise genetic distances can be used to quantify the amount of genetic ancestry that any admixed individual shows from the putative ancestral African, European and Native American

populations. The results of this kind of analysis are shown in Figure 2B and 2C. African ancestry represents the dominant admixture component for the vast majority of individuals from the population of Chocó analyzed here. The maximum fraction of African ancestry seen for any individual is 92.8%, and the average African ancestry for Chocó population is 75.8%. Nevertheless, there are substantial fractions of European and Native American ancestry seen for many of these individuals as well. The maximum fraction of European ancestry seen for any individ-

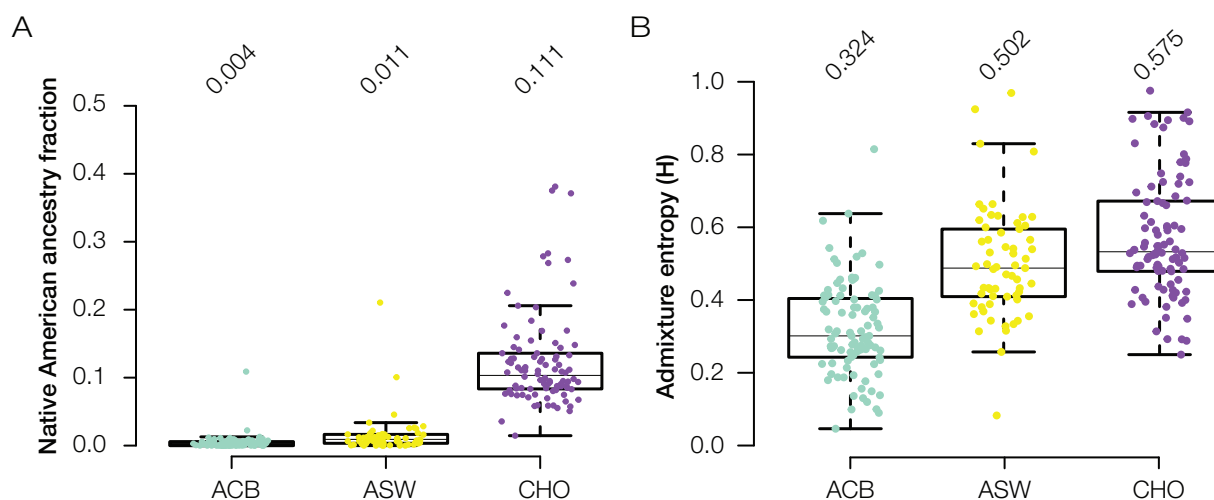


Figure 3. Distinct admixture characteristics of Chocó. Admixture patterns are compared among Chocó (CHO) and other New World African populations: ACB from Barbados in the Caribbean and ASW from the United States. (A) Native American ancestry and (B) admixture distributions for the three new world African populations; average values are shown for each distribution.

ual is 62.8%, and the average European ancestry is 13.1%. The maximum fraction of Native American ancestry is 39.6%, and the average Native American ancestry is 11.1%. The broad range of individual admixture percentages point to the diversity of the Chocó population.

The genetic ancestry of the population of Chocó shows some interesting differences compared to the genetic ancestry of the two other New World African populations analyzed here, the ACB population from Barbados in the Caribbean and the ASW population from the USA, despite the fact that all three populations show similarly high levels of overall African ancestry (~75-80%). First of all, the Chocó population (CHO) has substantially higher levels of Native American ancestry compared to the Afro-Caribbean (ACB) or African-American (ASW) populations (Figure 3A). Chocó has 11.1% average Native American ancestry, whereas the Afro-Caribbean and African-American populations have 0.4% and 1.1% average Native American ancestry, respectively. Second, the Chocó population shows higher levels of three-way genetic admixture, as measured by Admixture entropy (H), compared to the other two New World African populations (Figure 3B). This reflects the fact that in Chocó the non-African ancestry component is relatively evenly divided between European and Native American ancestry, whereas almost all non-African ancestry in the Caribbean and US populations is European. This pattern is indicative of longer and

more sustained contact between Afro-descendants and Indigenous communities in Chocó compared to what occurred in the Caribbean or the United States.

Sex-specific genetic ancestry. Mitochondrial DNA (mtDNA) and Y-DNA chromosomes are referred to as uniparental ancestry markers since they are inherited strictly along the maternal (for mtDNA) and paternal (Y-DNA) lineages. This means that mtDNA haplotypes can be used to infer female-specific ancestry and admixture, and Y-DNA haplotypes can be used to infer male-specific ancestry and admixture. The global origins of the mtDNA and Y-DNA haplotypes characterized from the population of Chocó show striking evidence of sex-specific ancestry in this population (Figure 4). The majority of mtDNA (82.1%) and Y-DNA (77.8%) haplotypes have African origins, consistent with the overall genetic ancestry of the population. However, the non-African ancestry components differ markedly for the female (mtDNA) versus male (Y-DNA) lineages. All of the non-African mtDNA haplotypes (17.9%) have Native American origins, whereas all of the non-African Y-DNA haplotypes have European (16.7%) or Middle Eastern (5.6%) origins. This sex-specific pattern of genetic ancestry may be linked to the unique historical conditions under which the state of Chocó was founded and populated (Wade 1995).

Genetic diversity levels in Chocó. As previously noted, based on its distinct population demographics, we propose that the state of Chocó is a rich source of

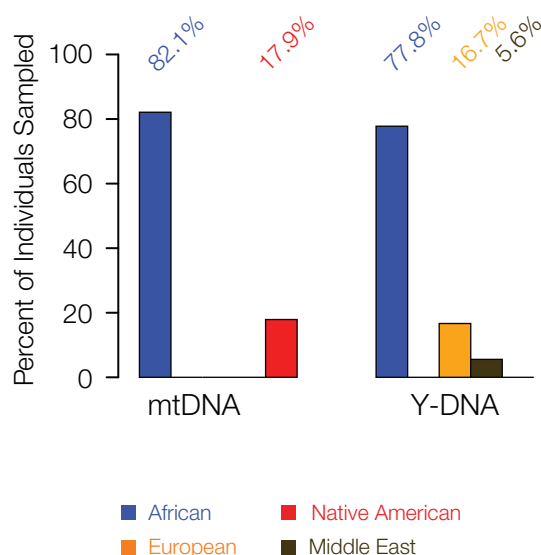


Figure 4. Sex-specific ancestry and admixture in Chocó. (A) Relative counts of African (blue), Native American (red), European (orange) and Middle Eastern (brown) mtDNA (maternal) and Y-DNA (paternal) haplotypes from the population of Chocó.

human biodiversity. To evaluate this proposition with respect to the genetic ancestry of the population, we compared the genetic diversity levels found in Chocó to levels of diversity seen for putative ancestral, non-admixed populations as well as other admixed American populations. The results of this analysis are shown in Figure 5. We evaluated genetic diversity in two ways: 1) via the overall scope of genetic distances between individuals in a population and 2) via the average genome-wide heterozygosity levels for all individuals in a population. The overall genetic diversity for each population was inferred by fitting a minimal spanning ellipse to the populations' pairwise genetic distance projection on the PCA plot (Figure 5A). The areas of the population-specific minimum spanning ellipses were then determined and used to quantify the populations' genetic diversity (Figure 5B). The Chocó population (CHO) has the highest level of overall genetic diversity calculated in this way for any of the populations analyzed here. The Chocó population also has high average heterozygosity levels compared to the other populations, second only to the other New World African population ASW (Figure 5C). Interestingly, the other admixed Colombian population from Medellín (CLM) also shows relatively high genetic diversity levels in these analyses despite the fact that it has a very different

genetic ancestry profile (*i.e.*, largely European ancestry) compared to the population of Chocó.

Conclusions

The ChocoGen collaborative research project has the joint aims of 1) characterizing the ancestry and genetic diversity of the population of Chocó, and 2) creating a genetic health profile of the population based on the diversity of its people. Investigators from UTCH and the Georgia Institute of Technology are collaborating to these ends, and this manuscript reports some of the first results of the project. The initial phase of the project is focused on ancestry analysis, and the results from this first phase will be used to inform the second health-related part of the effort. All of this work is being done in such a way as to develop the local human capacity for research in genetic ancestry and human health in the state of Chocó.

The analyses reported here indicate that the population of Chocó has an overwhelmingly African genetic ancestry, which is of course not surprising. Nevertheless, the Chocó population shows interesting differences, compared to other admixed American populations with similar levels of African ancestry, which likely reflect its distinct historical and cultural traditions (Wade 1995). In particular, individuals from Chocó show higher levels of three-way genetic admixture than other New World African populations, and this pattern can be largely attributed to the higher levels of Native American ancestry seen in Chocó. The population of Chocó also shows striking patterns of sex-specific ancestry, whereby non-African maternal ancestry is exclusively Native American, and non-African paternal ancestry is almost entirely European. This ancestry pattern may represent distinct admixture dynamics that characterized early (European admixture) from later (Native American admixture) historical periods in Chocó, and we plan to explore this idea further in subsequent studies.

Finally, the results on genetic ancestry and diversity obtained in this study underscore the extent to which Chocó represents a hotspot of human biodiversity. We hold that the human biodiversity of Chocó is an under-appreciated dimension of the area's well known biodiversity, and one that should be equally valued and fully developed for its potential.

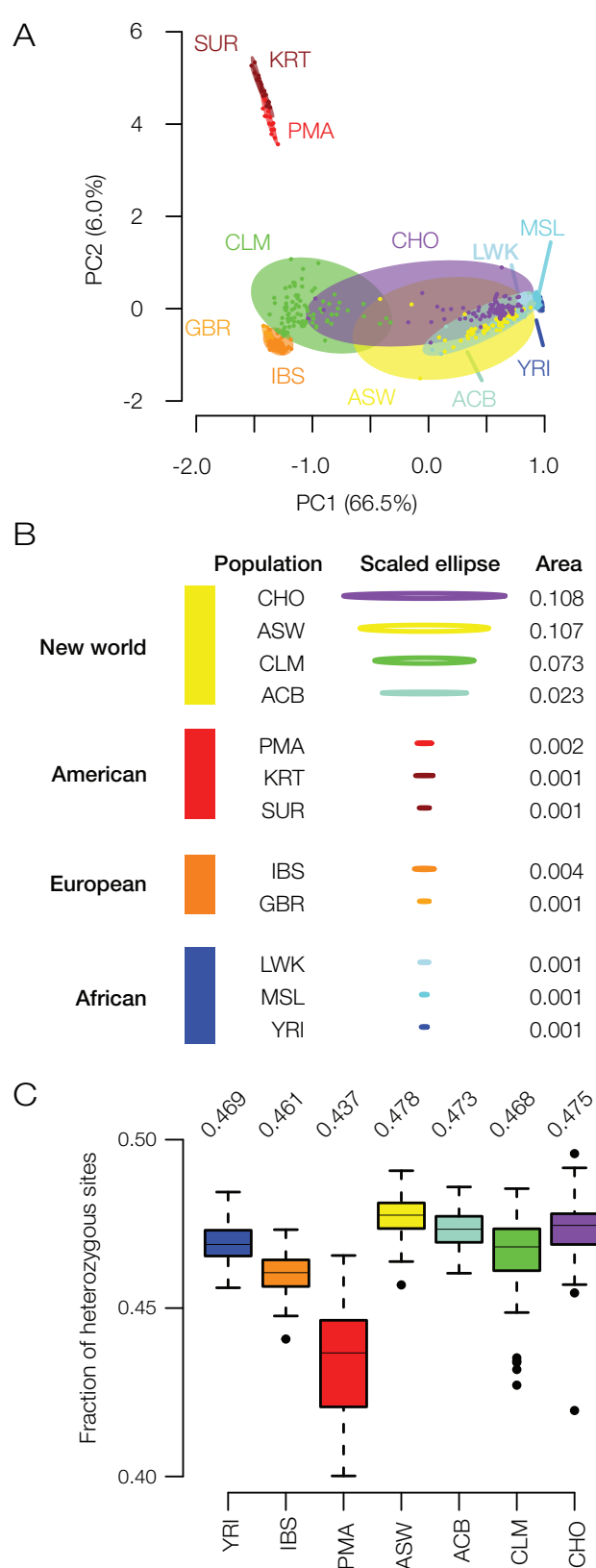


Figure 5. Genetic biodiversity in Chocó. (A) PCA plot showing the genetic relationships among individuals from the different human populations analyzed here (as shown and described for Figure 1), with each population bounded by a minimum spanning ellipse. (B) Scaled areas of the population-specific minimum spanning ellipses are used to quantify the overall genetic diversity of each population. (C) Distributions of the genome-wide heterozygosity values are shown for the populations analyzed here; average values are shown for each distribution.

Literature cited

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9): 1655-64.

Bedoya G, Montoya P, García J, Soto I, Bourgeois S, Carvajal L, *et al.* 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. *Proc Natl Acad Sci USA.* 103 (19): 7234-9.

Bryc K, Véléz C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, *et al.* 2010. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci USA* 107 (Suppl 2): 8954-61.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, *et al.* 2002. A human genome diversity cell line panel. *Science.* 296 (5566): 261-2.

Carvajal-Carmona LG, Ophoff R, Service S, Hartiala J, Molina J, Leon P, *et al.* 2003. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Hum Genet.* 112 (5-6): 534-41.

Carvajal-Carmona LG, Soto ID, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, *et al.* 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. *Am J Hum Genet.* 67 (5): 1287-95.

Córdoba L, García J, Hoyos LS, Duque C, Rojas W, Carvajal S, *et al.* 2012. Composicion genética de una población del suroccidente de Colombia. *Rev Colomb Antropol.* 48 (1): 21-48.

Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al.* 2010. A map of human genome variation from population-scale sequencing. *Nature.* 467 (7319): 1061-73.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, *et al.* 2015. A global reference for human genetic variation. *Nature.* 526 (7571): 68-74.

Hernández Romero A. 2005. *La visibilización estadística de los grupos étnicos colombianos.* Bogotá: Departamento Administrativo Nacional de Estadística (DANE).

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al.* 2002. The human genome browser at UCSC. *Genome Res.* 12 (6): 996-1006.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature.* 409 (6822): 860-921.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Rama-

- chandran S, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 319 (5866): 1100-4.
- Mann CC. 2013. *1493: Uncovering the new world Columbus created*. New York: Alfred A. Knopf.
- Markham C 1912. *The conquest of New Granada*. New York: EP Dutton and Company.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81 (3): 559-75.
- Rishishwar L, Conley AB, Vidakovic B, Jordan IK. 2015a. A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians. *Gene*. 574 (2): 345-51.
- Rishishwar L, Conley AB, Wigington CH, Wang L, Valderrama-Aguirre A, Jordan IK. 2015b. Ancestry, admixture and fitness in Colombian genomes. *Sci Rep*. 5: 12376.
- Ruiz-Linares A, Adhikari K, Acuna-Alonzo V, Quinto-Sánchez M, Jaramillo C, Arias W, et al. 2014. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet*. 10 (9): e1004572.
- Team RDC. 2008. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Wade P. 1995. *Blackness and race mixture: the dynamics of racial identity in Colombia*. Baltimore: JHU Press.
- Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, et al. 2008. Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4 (3): e1000037.
- Zachos FE, Habel JC. 2011. *Biodiversity hotspots: distribution and protection of conservation priority areas*. Vienna: Springer Science & Business Media.



UNIVERSIDAD DEL VALLE – FACULTAD DE SALUD
ESCUELA DE CIENCIAS BÁSICAS
POSTGRADO CIENCIAS BIOMÉDICAS.

ACTA DE SUSTENTACIÓN DE TESIS DOCTORAL
PROGRAMA ACADÉMICO DE DOCTORADO EN CIENCIAS BIOMÉDICAS (9695)

Jurado calificador conformado por los Doctores:

1. Dr. ADALBERTO SÁNCHEZ GÓMEZ
2. Dr. ARMANDO LUCUMÍ MORENO
3. Dr. JESUS FELIPE GARCIA VALLEJO
4. Dr. MIGUEL ANGEL MEDINA RIVAS

El día 15 de Julio de 2016, a las 9:00 a.m. se lleva a cabo la Sustentación de la Tesis Doctoral presentada por el estudiante de Doctorado en Ciencias Biomédicas (9695): Augusto Elías Valderrama Aguirre – 200306203.

Título de la Tesis Doctoral: “*Genómica de poblaciones colombianas: ancestralidad y adaptación*”



Dr. J. KING JORDAN

Director Tesis Doctoral



De acuerdo con los resultados anteriores se determina que la sustentación de la Tesis Doctoral presentada por el estudiante es:

(☒) Aprobada () Meritoria () Laureada () Reprobada

En caso de ser reprobada el estudiante debe acoger las recomendaciones del jurado y presentar nuevamente el documento ante el Director del Proyecto.

() Requiere () No requiere nueva sustentación.

El plazo para la nueva sustentación y/o para presentación del documento final es de:

Otros comentarios:

El trabajo sobre el porcelano de la
Investigación postdoctoral que se ha
colaborado y la importancia en la Salud
Humana de la población porque permite
información sobre las relaciones de Afectos
con marcadores biológicos de enfermedades
de importancia para la Salud pública
del país.



El Comité Gestor de la Universidad del Valle
que la Reglamentación de la Universidad
no permite dar la calificación de
Aprobado, lo que, si fuese
que debe ser personal como
los jurados

Dr. JULIO CÉSAR MONTOYA VILLEGAS
Director Postgrado en Ciencias Biomédicas

Jurado

Dr. ADALBERTO SÁNCHEZ GÓMEZ

Jurado

Dr. ARMANDO LUCUMÍ MORENO

Jurado

Dr. JESÚS FELIPE GARCÍA VALLEJO

Jurado

Dr. MIGUEL ÁNGEL MEDINA RIVAS

Santiago de Cali, Julio 14 de 2016

Señores

Comité de Postgrado

Escuela de Ciencias Básicas

Facultad de Salud

Ref: evaluación Tesis de doctorado Genómica de poblaciones Colombianas

Apreciados Señores:

En mi calidad de jurado de la tesis doctoral del estudiante Augusto Elías Valderrama Aguirre, Genómica de poblaciones colombianas: ancestralidad y adaptación, me permito remitir a ustedes el siguiente concepto:

El manuscrito cuenta con todas las condiciones de una tesis doctoral y excede los estándares de calidad y pertinencia de nuestro programa. El tema abordado es de la mayor relevancia y la utilización de metodología de análisis en el mismo demuestran una actualización acorde a exigencias internacionales de alto grado. La posibilidad de reconocer mecanismos de diferenciación y fijación alélica en la población colombiana tiene un alto valor desde el punto de vista clínico-epidemiológico con miras a la implementación de políticas de acción pública en la asignación de recursos para cubrir las demandas actuales. Por esta razón, este trabajo debe ser aprovechado como un punto de partida para el desarrollo de una línea temática nacional.

Por lo anterior expuesto, mi concepto como jurado es aprobatorio con mención "laureada".

Por último, solicito al Dr. Montoya, leer la siguiente pregunta en la ceremonia de sustentación:

Como asociar los resultados encontrados en este trabajo con el panorama actual epidemiológico de enfermedades infecciosas emergentes de tipo viral?

Mis felicitaciones al estudiante y su tutor.

Cordialmente

Adalberto Sanchez