Universidad de la República

Programa de Desarrollo de las Ciencias Básicas

Variabilidad Genómica y Ancestría en la población Mexicana: aplicación en estudios de asociación de enfermedades complejas.

T E S I S

para obtener el título de:

Doctor en Ciencias Biológicas

presenta:

Magíster Valentina Colistro

Tutora:

Dra. Mónica Sans

Co-Tutor:

Dr. Augusto Rojas-Martínez

Montevideo, Uruguay, 2021

Autor:
Mag. Valentina Colistro
Depto. de Métodos Cuantitativos
Facultad de Medicina - UdelaR

Directora:
Dra. Mónica Sans
Depto. de Antropología Biológica
Fac. de Humanidades y Cs. de la Educación - UdelaR

Co-director:
Dr. Augusto Rojas-Martínez
Líder del GIEE Genética Humana
Tecnológico de Monterrey - México

Tribunal:
Dr. Bernardo Bertoni
Depto. de Genética
Facultad de Medicina - UdelaR

Dra. Carolina Bonilla
Depto. de Medicina Preventiva
Faculdade de Medicina
Universidade de São Paulo

Dr. Julio da Luz
Laboratorio de Genética Molecular Humana
Depto. de Ciencias Biológicas
CENUR Litoral Norte-Salto - UdelaR

Cosas chicas para el mundo, pero grandes para mi...

Mi tapera - Elias Regules.

A mi padre.

Monstruos Mexicanos

En un país como México, muchos mundos se superponen bajo la tierra. Un universo subterráneo nutre nuestro territorio físico y espiritual. Basta con que escombremos la maleza de los montes para descubrir una pirámide. Basta con que excavemos un poco para encontrarnos las ruinas de un templo antiguo, la estatua de un dios olvidado o vasijas donde nuestros antepasados bebieron sus sueños. Basta con que escudriñemos en nuestros sueños para hallar una muchedumbre de criaturas imaginarias creadas por los ancestros. Y basta con que cavemos en cualquier sitio para encontrar una multitud de huesos enterrados.

También en nuestros huesos y en nuestra sangre sobreviven memorias de otros tiempos, de las que apenas nos damos cuenta en la vida diaria. Son memorias de pueblos que a lo largo de la historia imaginaron, trabajaron y lucharon entre si, y cuyo espíritu vive bajo nuestra piel.

Hace casi quinientos años, los conquistadores españoles llegaron con sus ideas y sus armas para imponerse sobre la gente que vivía en estas tierras. Cada grupo tenía sus ritos y dioses particulares -a menudo representados en la figura de una serpiente alada-, pero todo ello fue aplastado por el monstruo conquistador: un militar montado en su caballo, mitad hombre-mitad bestia, que traía un dios nuevo y disparaba sus arcabuces de pólvora como eructos de dragón. El dragón de fuego venido de Europa luchó, pues, con la serpiente sagrada americana y aparentemente la venció".

Carmen Leñero - Monstruos Mexicanos, 2012

Editorial Alas y Raíces

Agradecimientos

Le agradezco profundamente a todos los que estuvieron cerca durante este largo proceso, sería imposible nombrarlos a todos, aunque haré alguna excepción.

A Mónica, mi directora de tesis, quién constantemente me ayudó y motivó a superar los problemas a los que debí enfrentarme durante este proceso y a mi co-director Augusto Rojas-Martínez quien me recibió cálidamente y me ayudó a valorar el trabajo hecho.

A los organismos involucrados, PEDECIBA, UdelaR, CSIC y la CEE, todos soportes fundamentales para poder desarrollar esta tesis y también a las instituciones extranjeras que me recibieron: Universidad de Oxford, Universidad de Santiago de Compostela y al Tecnológico de Monterrey. También a todos los mexicanos que en pos de la ciencia cedieron sus datos para este proyecto.

A Alexandra Elbakyan, quién me dió acceso a la literatura científica que me premitó realizar esta investigación. A Bernardo Bertoni, por su crítica siempre constructiva.

A mi padre quien mucho me aportó desde su experiencia clínica a que yo pueda comprender las características del cáncer colorrectal. A mi madre por su apoyo incondicional y por mirada siempre optimista.

Por último pero no menos importante, un agradecimiento especial a Gastón, quien compartió día a día entusiasmos y frustraciones. Esta tesis también es trabajo de él.

Resumen

Este trabajo de tesis doctoral se titula: Variabilidad Genómica y Ancestría en la población Mexicana: aplicación en estudios de asociación de enfermedades complejas, y fue realizado bajo la tutoría de la Dra. Mónica Sans (UdelaR) y el Dr. Augusto Martínez-Rojas (Tecnológico de Monterrey, México).

La tesis aborda la temática del impacto que tiene la historia demográfica de mestizaje sobre la variabilidad genómica de la población de México y el efecto en la asociación de regiones genómicas a una enfermedad compleja, el cáncer colorrectal (CRC). Los objetivos consideran la determinación de proporciones de ancestría genética, estimaciones de parentesco utilizando diferentes abordajes metodológicos, determinar la asociación a CRC de regiones genómicas de una población mestizada y el impacto de estas estimaciones en los estudios genético poblacionales de poblaciones mestizadas.

Se analizaron 10 regiones genómicas previamente descritas como asociadas a CRC estimando la proporción de ancestría, el patrón de desequilibrio de ligamiento, la estructura haplotípica y los niveles de heterocigosidad. Los resultados evidencian diferencias entre las regiones cromosómicas y entre las poblaciones estudiadas, aunque cada región mostró sus particularidades. Se muestra la importancia de desarrollar modelos estadísticos que tengan en cuenta la historia demográfica de la población estudiada y las características particulares de las regiones cromosómicas.

A continuación se analizó la estructura poblacional de las muestras de casos y controles de México y cómo el mestizaje condiciona las estimaciones de parentesco entre las muestras. El trabajo mostró que se obtienen resultados espúreos al aplicar métodos no especìficamente diseñados para poblaciones mestizadas ya que se evidencia un sesgo en la ancestría nativo americana de las muestras marcadas como emparentadas. Sin embargo, al trabajar con modelos optimizados para poblaciones con subestructura genética se obtienen resultados marcadamente diferentes. Concluimos que se deben considerar modelos específicos que consideren la naturaleza mestizada de la población para dichas estimaciones.

Finalmente se analizaron datos genotípicos de las muestras de casos y controles de CRC de México en búsqueda de regiones genómicas asociadas a la enfermedad. El estudio se realizó con 831 casos y 881 controles que fueron genotipados para 1,006,703 polimorfismos autosómicos. El GWAS incluyó la estructuración poblacional y otras variables confusoras. Este análisis arrojó un SNP fuertemente asociado (rs35797542) en la región 8q24.22 y otros 16 SNPs con asociación sugerente (5×10^{-8}<valor-p $<1 \times 10^{-5}$). LA asociación genotípica se complementó con una prueba de asociación que considera el efecto combinado de todos los SNPs que yacen dentro de los límites de los genes asociados; Test de Asociación de Núcleo o SKAT (por sus siglas en inglés, Sequence-Kernel Association Test). Esta prueba detectó asociación en 5 genes (*LRRC36, PLEKHG4, KCTD19,ATP6V0D1 y ZFAT*), uno de ellos (*ZFAT*) es el gen donde está ubicado rs35797542.

Abreviaciones

1000G	Proyecto 1000 Genomas
ADN	Ácido desoxiribonucelico
AIM	Amcestry Informative Marker, Marcador Informativo de Ancestralidad
AMS	Admixture Mapping Study, Estudio de Mapeo por Mestizaje
APC	Adenomatous Polyposis Coli Protein
ARN	Ácido ribonucelico
ASW	Individuos Afroamericanos
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
CEU	Individuos descendientes de europeos
CHIBCHA	Genetic Study of Common Hereditary Bowel Cancers in Hispania and the Americas
ChrY	Cromosoma Y
CLM	Individuos de Colombia
CNV	Copy Number Variation, Variantes de número de copias

CONAPO	Consejo Nacional de Población
CRC	Cáncer Colorrectal
dbSNP	Base de Datos de SNPs
DL	Desequilibrio de ligamiento
eGWAS	EpiGenome-wide association Study, Estudio de asociación epigenómica
FDR	False Discovery Rate, Tasa de descubrimiento falso
GWAS	Genome-wide Association Study, Estudio de asociación de todo el genoma
GWRAS	Genome-wide Runs of heterocigocity Association Study, Estudio de asociación de todo el genoma de heterocigosidad continua.
HapMap	Proyecto HapMap
HCCS	Estudio Hispánico de Cáncer Colorrectal
HF-CRC	Historia Familiar de Cáncer Colorrectal
HWE	Equilibrio de Hardy-Weinberg
IBD	Identity-by-descent, Idénticos por descendencia
IC	Intervalo de confianza
IMC	Índice de Masa Corporal
LA	Latinoamérica
MAF	Frecuencia Alélica Mínima
MEX	Individuos mexicanos de CHIBCHA

MLH1	DNA Mismatch Repair Protein Mlh1
MSH2	DNA Mismatch Repair Protein Msh2
MSH6	DNA Mismatch Repair Protein Msh6
MT	Mitocondria
MUTYH	Adenine DNA Glycosylase
MXL	Individuos mexicanos de Los Ángeles
NAM	Individuos Nativo americanos
OMS	Organización Mundial de la Salud
OR	Odd ratio
PCA	Análisis de componentes principales
PEL	Individuos de Perú
PMS2	PMS1 Homolog 2, Mismatch Repair System Component
PS	Estructuración poblacional
PUR	Individuos de Puerto Rico
SKAT	Sequence Kernel Association Test, Prueba de asociación de núcleo
SNP	Single nucleotide polimorphism, Polimorfismos de un único nucleótido
UTR	Región no traducida
WHO	World Health Organization
YRI	Individuos Yoruba

Índice general

Desarrollo de la Tesis

Esta tesis se desprende del proyecto multicéntrico *"Genetic Study of Common Hereditary Bowel Cancer in Hispania and the Americas"* (CHIBCHA), coordinado por Ian Tomlinson y Luis Carvajal-Carmona del Departamento de Genética Poblacional y Molecular de *Wellcome Trust Centre for Human Genetics*, Universidad de Oxford, Reino Unido y financiado por el 7^{mo} Programa Marco de la Comunidad Económica Europea (Proyecto Nº:223678). También estuvieron involucrados centro de investigación de Colombia (Responsable: Maria Magdalena Etcheverry de la Universidad de Tolima), Brasil (Samuel Aguilar del Hospital do Cancer A.C. Camargo), México (Responsable: Augusto Rojas-Martínez de la Universidad Autónoma de Nuevo León), Portugal (Manuel Texeira del Instituto Portugués de Oncología de Porto), España (Responsable: Ángel Carracedo de la Universidad de Santiago de Compostela y Sergi Castellví-Bel de la Fundació Clinic per a la Recerca Biomèdica) y Uruguay (Responsable: Mónica Sans de la Universidad de la República). La tesis se desarrolla en cinco capítulos:

1. Introducción
2. Estructura poblacional
3. Análisis de ancestría en genes candidatos
4. Estudio de asociación de todo el genoma con cáncer colorectal (CRC)
5. Conclusiones generales y perspectivas

El capítulo 1 es una introducción, donde se desarrollan las generalidades y particularidades de los temas a considerar en los siguientes capítulos. Los capítulos 2, 3 y 4 se presentan en formato artículo de publicación.

El **capítulo 2** es un artículo publicado en *Genetics and Molecular Biology* en noviembre de 2020, con el título de ***"Differential admixture in Latin American populations and its impact on the study of colorectal cancer"*** (Colistro *et al.* 2020). En este caso se analizaron las proporciones de ancestría en 10 regiones genómicas previamente descritas como asociadas a CRC. El análisis lo realizamos considerando muestras de la población mexicana colectadas para el proyecto CHIBCHA y muestras del proyecto 1000G, disponibles públicamente. Además, analizamos el patrón de desequilibrio de ligamiento, la estructura haplotípica y los niveles de heterocigosidad. Los resultados evidencian diferencias entre las regiones cromosómicas y entre las poblaciones estudiadas (mestizadas y parentales) que nos permitieron concluir sobre los valores intermedios de variabilidad genómica de las poblaciones mestizadas respecto a las parentales aunque cada región mostró sus particularidades, que deben ser tenidas en cuenta al momento de determinar asociación. El estudio muestra la importancia de desarrollar modelos estadísticos que tengan en cuenta la historia demográfica de la población estudiada y las características particulares de las regiones cromosómicas.

El **capítulo 3** es una Short Communication aceptado en la revista *Annals of Human Genetics*, titulado ***"Population structure and relatedness estimates in a Mexican sample"*** (Colistro et. al 2021). En este capítulo se analiza la estructura poblacional de las muestras de casos y controles de México y cómo el mestizaje condiciona las estimaciones de parentesco entre las muestras. La estructura poblacional ha sido ampliamente estudiada como factor confusor en los Estudios de Asociación de Genoma Completo (GWAS) y las poblaciones Latinoamericanas ofrecen un escenario idóneo para estos estudios. En este trabajo analizamos cómo influye el mestizaje en las estimaciones parentesco entre las muestras usando como indicador las estimaciones de idénticos por descendencia, IBD por su siglas en inglés (identity-by-descent).

En el trabajo mostramos cómo se obtienen resultados espúreos al aplicar métodos no especìficamente diseñados para poblaciones mestizadas ya que se evidencia un sesgo en la ancestría nativo americana de las muestras marcadas como emparentadas. Sin embargo, al trabajar con modelos optimizados para poblaciones con subestructura genética se obtienen resultados marcadamente diferentes y en sintonía con la realidad poblacional de la muestra estudiada, la población mexicana. Concluimos que no es correcto aplicar modelos estadísticos diseñados para poblaciones homogéneas desde el punto de vista de ancestría genética, sino que se deben considerar modelos específicos que consideren la naturaleza mestizada de la población.

El **capítulo 4** tiene el título de ***"A genome-wide association study of colorectal cancer in Mexican mestizos suggest novel common tumor-risk variants"*** y se presenta la versión a ser enviada para su evaluación en *Frontiers in Oncology*. Analizamos datos genotípicos de las muestras de casos y controles de CRC mexicanas en búsqueda de regiones genómicas asociadas a la enfermedad. El estudio se realizó con 831 casos y 881 controles que fueron genotipados para 1,006,703 polimorfismos autosómicos. El GWAS se realizó mediante la aplicación de un modelo de regresión que incluyó la estructuración poblacional y otras variables que han sido reportadas como variables confusoras de la asociación con CRC. Este análisis arrojó un SNP fuertemente asociado (*rs35797542*; valor p $< 5 \times 10^{-8}$), localizado en la región 8q24.22 y otros 16 SNPs mostraron una asociación sugerente ($5 \times 10^{-8} <$ valor-p $< 1 \times 10^{-5}$). Para confirmar los resultados de asociación genotípica realizamos una prueba de asociación que considera el efecto combinado de todos los SNPs que yacen dentro de los límites de los genes asociados; Test de Asociación de Núcleo o SKAT (por sus siglas en inglés, Sequence-Kernel Association Test). BienEsta prueba detectó asociación en 5 genes (*LRRC36, PLEKHG4, KCTD19, ATP6V0D1 y ZFAT*), uno de ellos (*ZFAT*) es el gen donde está ubicado rs35797542. Este *locus* no ha sido previamente reportado como asociado a CRC, y contribuye al conocimiento de la etiología genética del CRC aunque futuros estudios en poblaciones latinoamericanas serán necesarios para replicar los resultados; asimismo, los restantes 4 genes también son potenciales candidatos que merecen ser estudiados en mayor profundidad. Una vez más el estudio muestra la importancia

de realizar GWAS en poblaciones mestizadas.

El **capítulo 5** contempla integralmente conclusiones de los resultados obtenidos en los capítulos anteriores teniendo en cuenta que las conclusiones específicas de cada tema puntual se exponen en los respectivos capítulos.

Capítulo 1

Introducción

1.1. Cáncer y cáncer colorrectal

1.1.1. Epidemiología

El cáncer es uno de los problemas de salud pública más grande a nivel mundial del siglo XXI, con un crecimiento sostenido en la mayoría de los continentes. Datos de la Organización Mundial de la Salud (OMS) lo posicionan como la 2^{da} causa de muerte en todo el mundo, siendo que el 1^{er} lugar lo ocupan las enfermedades cardiovasculares. En el año 2018, el proyecto Globocan, estimó que hubo a nivel mundial 18.1 millones de nuevos casos de cáncer, 9.5 millones de muertes por cáncer y 32.6 millones de personas viviendo con esta enfermedad. Dentro de los cánceres, el de pulmón lidera las causas de muerte a nivel mundial, con casi 2 millones de muertes en 2019, seguido en segundo lugar por el CRC con casi 1 millón de muertes el mismo año. Al observar los datos por sexo, el CRC pasa a 3^{er} lugar tanto en hombres como en mujeres. En los hombres es superado por cáncer de pulmón y de hígado y en las mujeres, por cáncer de mama y de pulmón (Global Cancer Observatory, 2019).

Además, una consideración a tener en cuenta respecto al cáncer en el continente americano, es que a pesar de la creciente incidencia, el cáncer se ve ensombrecido por las enfermedades transmisibles (según la Sociedad Americana del Cáncer http://www.cancer.org), lo cual genera una ausencia de campañas eficaces para la sensibilización y la detección temprana, generando que el cáncer se detecte en las últimas etapas, cuando el tratamiento es menos eficaz. Esta situación se agrava en países de bajos o medianos ingresos ya que las enfermedades transmisibles ocupan gran parte de la atención y provoca que se descuiden u olviden campañas de prevención ("WHO | Noncommunicable Diseases: The Slow Motion Disaster," 2017).

Los países de América Latina y el Caribe también presentan tasas de incidencia y prevalencia de cáncer crecientes. En los últimos 30 años la tasa de mortalidad por cáncer aumentó de 71.6 muertes cada 100,000 habitantes en 1990 a 109.2 muertes cada 100,000 habitantes en 2019. Globocan estima que en 2019 murieron 650,000 personas en el continente por cáncer. Al igual que en el resto del mundo, es la 2^{da} causa de muerte luego de las enfermedades cardiovasculares. En estos países los tres cánceres que más muertes causan son pulmón, CRC y mama, entre los tres fueron responsables en 2019 de 212,000 muertes. Al discriminar por sexos el CRC ocupa el 3^{er} lugar tanto en hombres como en mujeres. En el caso de los hombres es precedido por próstata y pulmón, y en las mujeres por mama y pulmón. En 2019, el CRC causó 10.9 muertes por cada 100,000 habitantes en América Latina y el Caribe (Global Cancer Observatory, 2019).

En Latinoamérica, a pesar de presentar tasas de prevalencia menores a las europeas, las tasas de mortalidad e incidencia de CRC han crecido en las últimas tres décadas. En la década de 1990 el CRC ocupaba el 13^{er} puesto entre las causas de muerte (14.59 muertes cada 100,000 habitantes) mientras que en 2010 ocupaba el 8^{vo} puesto (19.05 muertes cada 100,000 habitantes). El CRC presenta variación en los valores de incidencia según la región geográfica, encontrándose las tasas más altas en países de medianos y altos ingresos (55 % de los casos del mundo). La tasa de incidencia más alta en 2019 se observó en Australia y Nueva Zelanda, seguido por los países europeos, mientras que las tasas más bajas pertenecen a África,

según datos del Institute of Health Metrics and Evaluation de la Universidad de Washington (http://www.healthmetricsandevaluation.org/).

En el caso de México, las neoplasias ocupan desde el año 2000, el 3^{er} puesto de mortalidad por debajo de las enfermedades cardiovasculares y la diabetes en ambos sexos, con 85.2 muertes cada 100,000 habitantes en 2019, posición que se mantiene al analizar hombres y mujeres por separado. Dentro de los cánceres el CRC ocupa el 2^{do} lugar luego del cáncer de pulmón (8.4 muertes cada 100,000 habitantes en 2019) y esta posición cambia al 3^{er} puesto al considerar por separado los dos sexos, siendo precedido por cáncer de mama y de cérvix en las mujeres y en los hombres por próstata y pulmón (Data Visualizations | Institute for Health Metrics and Evaluation, 2020).

1.1.2. Factores de Riesgo

El CRC no tiene grandes factores de riesgo identificados, sino que sigue un modelo de muchos factores de bajo riesgo y sus interacciones (Martínez, 2005). Dentro de los factores no modificables se destaca la edad, los antecedentes familiares de CRC, la enfermedad inflamatoria intestinal y los factores genéticos; respecto a los factores modificables, la dieta, la obesidad, el tabaquismo y el consumo de alcohol son los más frecuentemente asociados a CRC. A pesar de tener un efecto bajo o moderado, el efecto de cambiar estos hábitos no está claro aún.

Factores de riesgo no modificables

En el caso de la edad, se ha observado un aumento del riesgo en personas mayores de 50 años, donde la incidencia se duplica con cada década (Farreas-Rozman, 2013), antes de los 40 años es muy bajo, siendo prácticamente todos los casos diagnosticados a esta edad atribuibles a antecedentes familiares de CRC (Boyle & Leon, 2002). Los datos de Globocan muestran que

a nivel mundial el CRC tiene una tasa de incidencia de 3.2 casos cada 100,000 habitantes en personas menores de 50 años mientras que para personas entre 50 y 70 años la tasa es de 62.1 casos cada 100,000 habitantes para ambos sexos, siendo más alta en hombre que en mujeres en ambos grupos etarios.

Otro factor de riesgo que ha sido descrito es el de los antecedentes familiares de CRC, ya que la presencia de historia familiar con CRC (HF-CRC) aumenta el riesgo a presentar la patología en el individuo (Henrikson *et al.*, 2015). Estos autores hicieron una revisión sistemática en la cual se analizaron 9 artículos de diversos países y con tamaños poblacionales variados que midieron el riesgo a CRC en individuos con y sin HF-CRC, y encontraron que en 8 de los 9 artículos el riesgo fue significativamente mayor en presencia de HF-CRC.

Por último la presencia de enfermedad inflamatoria intestinal muestra un mayor riesgo a CRC en los individuos que padecen esta patología, aunque esta condición contribuye minoritariamente a la incidencia total de CRC (Boyle & Leon, 2002).

Factores de riesgo modificables

Dieta

Un metaanálisis realizado en 2011 (Chan *et al.*, 2011) que examina 28 estudios de cohorte referidos al consumo de carnes rojas y su correlación con la incidencia de CRC muestran resultados que indican que consumir más de 100 g diarios de carne roja aumenta un 17 % el riesgo a CRC. Sin embargo estos valores son muy dispares entre las poblaciones de diferentes continentes y en el caso de las poblaciones asiáticas, al ser analizadas por separado, no muestran un aumento significativo del riesgo relativo al CRC.

En consonancia, una revisión reciente (Hur *et al.*, 2019) puso el foco en los alimentos ricos en hierro hemínico y arroja resultados que hacen difícil concluir que las carnes rojas sean

las principales responsables de mayor riesgo de CRC, ya que por sí solas no son la principal fuente de hierro hemínico, y que depende del método de preparación, condimentos agregados y alimentos acompañantes.

Obesidad

Una revisión sobre índice de masa corporal (IMC) y cáncer, publicada en 2008, indica que en personas obesas (IMC $\geq$ 30 kg/m^2) el riesgo de padecer CRC es entre un 1,5 y 3,5 mayor que entre los que tienen un IMC menor (Pischon *et al.*, 2008). La relación es más evidente en determinados tumores entre los cuales se mencionan CRC, cáncer de mama en mujeres postmenopáusicas, páncreas, próstata e hígado, así como también en menores de 55 años de ambos sexos.

Un análisis más reciente y específico de CRC y obesidad se centra no solo en la epidemiología mundial de ambas condiciones, sino también en los mecanismos moleculares que podrían explicar el vínculo entre ellas (Ye *et al.*, 2020). Concluye que la correlación epidemiológica es clara y ampliamente estudiada y que los mecanismos moleculares subyacentes a la obesidad son consistentes con los ambientes celulares que propician el crecimiento de células tumorales.

Tabaquismo y alcohol

Diversos estudios han reportado un aumento del riesgo a CRC en fumadores activos: un metaanálisis del año 2008 muestra resultados que indican que la condición de fumador es determinante tanto en la formación como en la severidad del CRC (Botteri *et al.*, 2008). Este meta-análisis se basó en 42 estudios y reporta un efecto combinado de un riesgo de 2.14 ($IC_{95\,\%}$: 1.9 – 2.5), valor que decrece a 1.8 ($IC_{95\,\%}$: 1.7 – 2.0) si se analizan por separado fumadores actuales de ex-fumadores. Sin embargo, detectan un sesgo de publicación que lo explican por el período de latencia entre la exposición del humo de tabaco y la aparición de CRC. Está ultima consideración deja en evidencia la necesidad de contar con estudios de cohorte con tiempos de

seguimiento más prolongados y de estudios funcionales para poder determinar causalidad entre CRC y el tabaco. Otro estudio (Hannan *et al.*, 2009) determina que la exposición a humo de tabaco muestra asociación con el CRC únicamente si el hábito se mantiene durante un largo período y también evidencia la necesidad de estudios con seguimientos prolongados.

Respecto al alcohol, la evidencia es más difusa. Muchos estudios sobre estilo de vida y patrones de consumo muestran que la ingesta de alcohol es una variable de ajuste significativa para explicar el CRC.

Por último, un estudio publicado recientemente el cual analizó variables relacionadas al estilo de vida en 11,000 controles y 3895 casos de CRC mostró que fumar y consumir alcohol están asociadas significativamente a CRC (Hannan *et al.*, 2009). Por otra parte, el consumo de aspirina y otros antiinflamatorios no esteroides mostraron efectos auspiciosos en la reducción del riesgo a CRC (Flossmann & Rothwell, 2007).

1.1.3. Factores Genéticos

Los factores genéticos del CRC han sido objeto de estudio y se estima que son responsables del 35 % de los casos (Lichtenstein *et al.*, 2000), sin embargo solo un 5 % de los casos son atribuibles a mutaciones genéticas de alta penetrancia (Li, 1995). En la Tabla 1.1 se listan los genes identificados como de alta penetrancia en la población europea.

Si bien las mutaciones en estos genes explicarían un bajo porcentaje de los casos, al considerar el efecto combinado de múltiples mutaciones de bajo o moderado efecto, el componente genético cobra relevancia. Algunos estudios realizados en gemelos de poblaciones nórdicas detectaron que el efecto de los factores genéticos en el CRC se ubica entre el 27 %-42 % (Ahlbom *et al.*, 1997; Lichtenstein *et al.*, 2000; Verkasalo *et al.*, 1999), aunque todos los autores advierten que los resultados deben ser interpretados a la luz de la población estudiada.

Tabla 1.1: Genes de alta penetrancia asociados a CRC

Gen	Referencia	Cromosoma	Citobanda	ID Ensembl	Longitud (bp)
APC	Fodde, 2002	5	q22.2	ENSG00000134982	138,741
BRAF	Fransén *et al.*, 2004	7	q34	ENSG00000157764	205,602
MLH1	Dowty *et al.*, 2013	3	p22.2	ENSG00000076242	57,496
MSH2	Dowty *et al.*, 2013	2	p21	ENSG00000095002	260,079
MSH6	Klarskov *et al.*, 2011	2	p16.3	ENSG00000116062	114,533
MUTYH	Zorcolo *et al.*, 2011	1	p34.1	ENSG00000132781	11,730
PMS2	Broeke *et al.*, 2018	7	p22.1	ENSG00000122512	38,181

Un enfoque abarcativo, multivariable, que tenga en cuenta el riesgo conjunto asociado a múltiples mutaciones genéticas podría mejorar significativamente este porcentaje pero para esto es necesario previamente conocer no solo las variantes genéticas de alta penetrancia sino también las de efecto bajo o moderado específicas de cada población. De esta manera, la información genotípica nos permitiría establecer el riesgo de cada individuos a desarrollar CRC.

La principal razón para estudiar e identificar todos los polimorfismos causales de la enfermedad radica en poder generar intervenciones terapéuticas, pero también, para poder trabajar a nivel de la prevención ya que de contar con un panel exhaustivo que permita estimar el riesgo individual a CRC, se podrá ofrecer una vigilancia más intensiva a las personas con mayor riesgo; por ejemplo, en algunos países se ofrece una colonoscopia exploratoria en programas de prevención en salud a personas con HF-CRC (Levin *et al.*, 2008). Estrategias similares son utilizadas en otros cánceres tal como el de mama (Pharoah *et al.*, 2008).

Los métodos de identificación de polimorfismos asociados a CRC, para una subsecuente exploración de causalidad, han ido evolucionando a un ritmo vertiginoso en las últimas décadas. Desde comienzos del 2000 a la fecha, los GWAS cobraron central importancia en este aspecto.

1.2. Estudios de Asociación de todo el genoma (GWAS)

La investigación en genética humana creció exponencialmente en la segunda mitad del siglo XX. Los avances en la detección de mutaciones, modos de herencia, mapas de ligamiento, endogamia, estructura poblacional y la asociación de determinados polimorfismos con enfermedades, en conjunto con el desarrollo de las herramientas informática permitieron análisis a gran escala en constante evolución. Con el paso del tiempo, aumentaron los estudios a nivel poblacional y esto contribuyó a generar bases de datos que documentan la variabilidad intra e inter poblacional donde una de las formas de estimarla es considerar la variabilidad de frecuencias de las mutaciones.

En 2001 y respondiendo a la necesidad creciente de disponibilizar y sistematizar la información de la variabilidad del genoma humano, se publica la base de datos de dbSNP (Sherry *et al.*, 2001) cuyo crecimiento ha acompasado el aumento de datos generados en las dos últimas décadas. Casi simultáneamente en 2003 la culminación del Proyecto Genoma Humano (The Human Genome Project, 2020) y en 2005 la aparición del Proyecto HapMap (International HapMap Project, 2020) potenciaron las posibilidades de análisis poblacionales de gran escala. Estos esfuerzos en su conjunto marcan el punto de inflexión desde el cual los investigadores cuentan con herramientas que incluyen bases de datos computarizadas que contienen la secuencia del genoma humano de referencia, un mapa de la variación genética humana y un conjunto de nuevas tecnologías que pueden analizar de forma rápida y precisa muestras de genoma completo. Al día de hoy, dbSNP cuenta con más de 670 millones de variantes cortas registradas de los cromosomas autosómicos, sexuales y ADN mitocondrial. DbSNP no es la única plataforma disponible, pero la mayoría de las otras bases de datos sincronizan su información con ella.

Es en este contexto que cobra pertinencia y relevancia realizar estudios que consideren la variabilidad genética entre personas con y sin determinado rasgo para poder detectar diferencias estadísticamente significativas que den indicios de regiones genómicas asociadas al rasgo. Los

GWAS consisten en escanear una gran cantidad de marcadores polimórficos distribuidos a lo largo de todo el genoma de muchas personas para encontrar variaciones genéticas asociadas con una determinada enfermedad. Si bien los polimorfismos más ampliamente utilizados en los GWAS son los de un único nucleótido (SNP), no son los únicos. También se han realizado GWAS de polimorfismos respecto al número de copias, CNV, por su siglas en inglés (*Copy Number Variation*) (Craddock *et al.*, 2010) y también se han publicado versiones de GWAS que analizan el epigenoma (eGWAS) (Rakyan *et al.*, 2011). Además, los GWAS no son exclusivos para genomas humanos, también fueron realizados en una gran variedad de especies vegetales (Carlson *et al.*, 2019; Huang *et al.*, 2020; Khlestkin *et al.*, 2020) y animales (Kudinov *et al.*, 2019; Liu *et al.*, 2019; Xue *et al.*, 2020). En el caso de los animales, más específicamente del ganado, se ha desarrollado una versión de GWAS que estudia secuencias de homocigosidad ininterrumpidas a lo largo del genoma, denominada GWRAS (*Genome-wide run-of-homozygosity association study*) (Cesarani *et al.*, 2020).

Independientemente del polimorfismo y del rasgo considerado, el procedimiento consiste en examinar el genoma de cada participante en busca de polimorfismos genéticos seleccionados que muestren distribuciones alélicas significativamente diferentes en las personas con la enfermedad (”casos”) en comparación con las personas sin la enfermedad (“controles”); si se identifican, se dice que el polimorfismo está asociado con la enfermedad, aunque estos no necesariamente son los causante de la enfermedad, sino que son candidatos a seguir siendo estudiados en el contexto genómico que se encuentran.

En las últimas dos décadas un número creciente de GWAS se han desarrollado (Figura 1.1) y a medida que los avances tecnológicos permitan reducir costos y tiempos de análisis así como aumentar los polimorfismos a estudiar, sería de esperar que estos estudios sigan aumentando. Los rasgos más estudiados en humanos han sido el IMC, Diabetes Tipo II, asma e hipertensión arterial; a su vez, entre las neoplasias más estudiadas figura en primer lugar el cáncer de próstata seguido por cáncer de mama, de cérvix y CRC (GWAS Catalog, 2020).

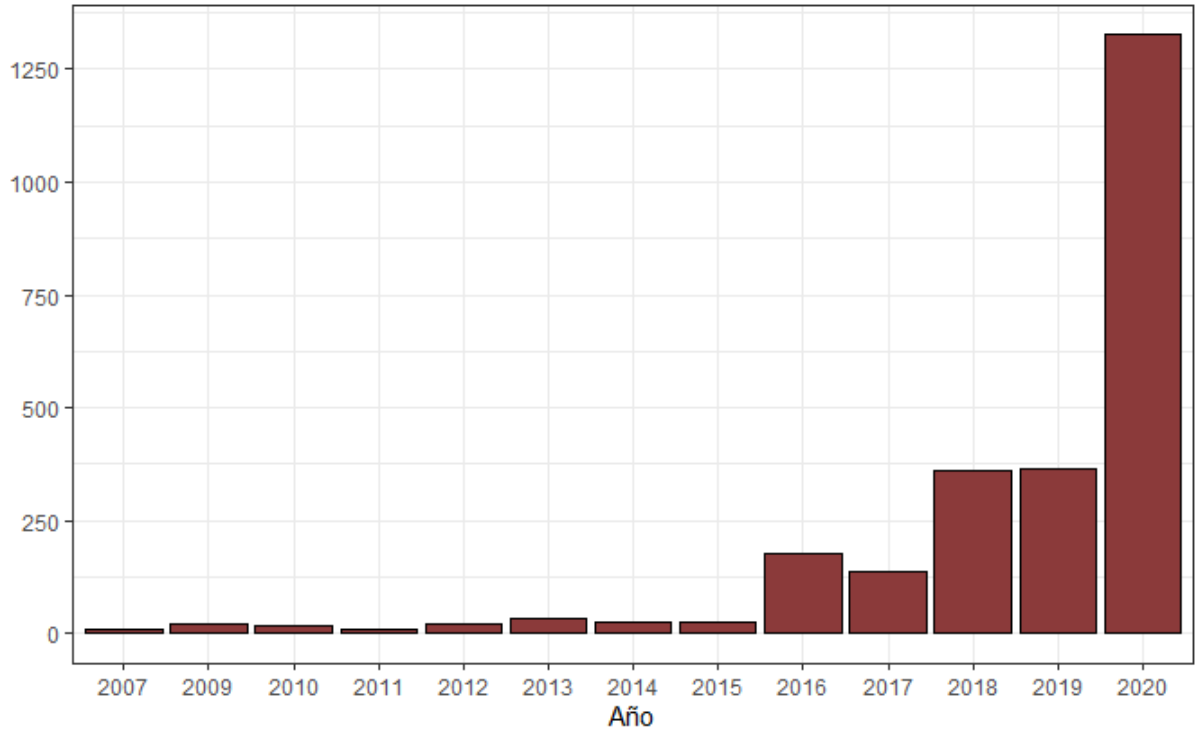

Figura 1.1: Cantidad de GWAS publicados desde 2007 hasta diciembre de 2020 discriminado por año de publicación. Datos obtenidos de la plataforma GWAS Catalog del EBI (https://www.ebi.ac.uk/gwas/home).

1.2.1. Estudios de asociación en CRC

El CRC también ha sido objeto de estudios genómicos en busca de regiones asociadas, varios GWAS han identificado *loci* de riesgo. Entre 2000 y 2010 una serie de GWAS en CRC identificaron nuevos *loci* asociados (Broderick *et al.*, 2007; Carvajal-Carmona *et al.*, 2011; Fernandez-Rozadilla *et al.*, 2013; Haiman *et al.*, 2007; Houlston *et al.*, 2008; Jaeger *et al.*, 2008; Tenesa *et al.*, 2008; Tomlinson *et al.*, 2007, 2008; Zanke *et al.*, 2007); todos estos estudios consideran muestras de individuos con ascendencia europea. Si bien estas investigaciones tuvieron variaciones metodológicas respecto al diseño de estudio, tamaño muestral, plataforma de genotipado utilizada y tratamiento de los datos, la uniformidad en la población estudiada permitió realizar un análisis conjunto de los datos combinados de todos los GWAS mencionados. Es así

que en 2013 se publicó un metaanálisis que considera el conjunto de los GWAS publicados en CRC a la fecha (Peters *et al.*, 2013). Al considerar los resultados obtenidos en los estudios individuales y en el metaanálisis se observó una baja reproducibilidad.

No pasó mucho tiempo para que se publicaran GWAS en CRC en poblaciones diferentes a la europea. Los primeros estudios en no europeos se realizaron en asiáticos (Cui *et al.*, 2011) y otros específicamente en poblaciones del sudeste asiático (Jia *et al.*, 2013; Zeng *et al.*, 2016; Zhang *et al.*, 2014). La incorporación de poblaciones diversas confirmó la baja reproducibilidad de los resultados de los GWAS y esto favoreció el desarrollo de estudios con muestras de múltiples orígenes dando lugar a estudios multiétnicos. Entre ellos un estudio que combina muestras asiáticas con muestras africanas detecta asociación en dos SNPs no previamente descritos, pero cuando intentan replicar los resultados en un metaanálisis en muestras europeas (16,000 casos y 18,000 controles) solo en uno de los SNPs replica la asociación (Wang *et al.*, 2014).

En el año 2016, se publica el primer GWAS de CRC que incluye individuos latinoamericanos (Schmit *et al.*, 2016). Este estudio contó con muestras provenientes de dos estudios, por un lado una cohorte que incluye 22 % de individuos latinos (Kolonel *et al.*, 2000), y por otro lado, una muestra del estudio HCCS (Hispanic Colorectal Cancer Study), de individuos auto identificados como hispanos de Estados Unidos de América y con diagnóstico confirmado de CRC. El tamaño muestral es significativamente menor a los estudios precedentes pero es la primera que incorpora una población mestizada. Este estudio no detecta ningún SNPs significativo.

Las poblaciones afroamericanas tampoco estuvieron ajenas a estos estudios y en 2017 se publicó un GWAS de CRC realizado exclusivamente en individuos afroamericanos (Wang *et al.*, 2017). En esta oportunidad se detectó una variante en un *locus* no previamente identificado (rs56848936), se replicó una variante (rs10411210) identificada en europeos (Houlston *et al.*, 2008) y otra variante (rs7252505), previamente asociada a CRC en otro estudio que incluía afroamericanos (Wang *et al.*, 2013), mostró valores de asociación muy cercanos a la significación

estadística.

Estos estudios y la no linealidad en la replicación de los resultados, evidencian la importancia de considerar estudios en poblaciones ancestralmente diversas y alerta sobre la extrapolación de los resultados de estudios entre poblaciones étnica e históricamente distantes. Es en este contexto que la composición ancestral de la población estudiada cobra importancia, en particular al trabajar con poblaciones mestizadas ya que son una excelente oportunidad de encontrar nuevas variantes que contribuyan a reducir el alto porcentaje de heredabilidad *ausente*. En particular Schubert *et al.* 2020 analizan las razones de la heredabilidad *ausente* del CRC y entre otras propuestas sugieren que se necesita más investigaciones en poblaciones étnicamente diversas para poder identificar nuevos alelos de susceptibilidad.

1.3. Las poblaciones Latinoamericanas

La población de América Latina y el Caribe representa el 6 % de la población mundial, aproximadamente 650 millones de personas (World Population Prospects - Population Division - United Nations, 2020). La realidad de cada región es sumamente diversa, pero hay un común denominador: su historia sociodemográfica. La población de América Latina es el resultado de un proceso de mestizaje de fundamentalmente tres poblaciones que podemos considerar como "parentales": nativo-americanos, europeos y africanos. Las fechas iniciales de los contactos entre los grupos se estiman hacia finales del siglo XV entre los dos primeros y a principios del siglo XVI entre los africanos y los otros dos grupos. Luego del contacto inicial el flujo de individuos hacia América Latina no ha cesado a lo largo del tiempo, tal como identifica Sans (2000).

Entre las consecuencias de la llegada de los europeos al continente se destacan los actos bélicos y las enfermedades infecciosas, ambas provocaron una disminución del tamaño poblacional de los pobladores originarios; asimismo, ya desde el inicio de la conquista, y facilitado por la

falta de mujeres europeas y por necesidades de alianzas estratégicas con los indígenas, comenzó el crecimiento de grupos mestizados. Este proceso se extendió por todo el continente aunque en cada región las particularidades histórico- demográficas generaron variaciones particulares que tuvieron efecto sobre las proporciones de mezcla entre pobladores originarios, europeos y africanos. Estas variaciones son observables en las poblaciones latinoamericanas actuales tanto a nivel social, cultural, lingüístico, político y genético, como resumen Mörner (1967) y Ruiz-Linares *et al.* (2014).

En lo relacionado a análisis genéticos poblacionales, se han identificado en el contexto actual el aporte de las tres (o, en algunos lugares, cuatro) poblaciones parentales. Este aporte ha sido ampliamente estudiado a nivel del cromosoma Y, ADN mitocondrial y autosómico en varios países latinoamericanos. Poblaciones actuales de las Antillas, Caribe y nordeste de Brasil, que fueron los primeros puertos de entrada de esclavos y que vieron diezmadas sus poblaciones indígenas, tienen una mayor proporción africana. Oleadas más recientes de migrantes europeos han cambiado el acervo genético de países como Argentina, Uruguay y el sur de Brasil, donde los habitantes tienen los niveles más altos de ascendencia europea en Latinoamericana (Salzano & Sans, 2014).

En Uruguay se evidenció el aporte nativo tanto en estudios de ADN mitocondrial, nuclear y cromosoma Y (Bertoni *et al.*, 2005; Bonilla *et al.*, 2004, 2015; Da Luz *et al.*, 2010; Gascue *et al.*, 2005; Hidalgo *et al.*, 2005; Sans *et al.*, 1997, 2002; entre otros); en Argentina la composición ancestral es muy diversa a lo largo del país (Fejerman *et al.*, 2005; Goicoechea *et al.*, 2001; Martínez Marignac *et al.*, 2004; entre otros), Brasil presenta el aporte africano mayor del continente (ver revisión, Salzano & Sans, 2014), países andinos como Perú muestran un gran componente nativo (Sandoval *et al.*, 2013), Colombia por su parte ha demostrado tener niveles bajos de ancestría africana con excepciones de ciertas regiones (Rojas *et al.*, 2010) y en Guatemala (Martínez-González *et al.*, 2012) se observa un gran aporte por parte de la parental nativa-americana. México también presenta una historia de mestizaje con el aporte de tres

parentales y nuevamente se observa un mayor aporte nativo-americano así como una diferencia entre marcadores del cromosoma Y y del ADN mitocondrial que denota uniones direccionales entre hombres europeos y mujeres indígenas, aportes que serán detallados en el siguiente capítulo.

Todos estos estudios muestran que el cambio que generó el proceso de mezcla poblacional en la estructura genética de estas poblaciones las cuales se han evaluado mediante análisis de todo el genoma tanto en poblaciones mexicanas como latinoamericanas, a los efectos de este estudio nos centraremos en la población actual de México.

1.3.1. Población Mexicana

En México residen 124.9 millones de personas (Datos de la Encuesta Nacional de la Dinámica Demográfica, 2018 https://www.inegi.org.mx/programas/enadid/2018/). Esta población es el resultado de varios eventos que han tenido lugar desde la conquista europea hasta nuestros días. En México, la colonización europea y la introducción de poblaciones esclavizadas provenientes de África dio lugar a un complejo proceso de mezcla biológica principalmente entre nativos americanos, españoles y esclavos africanos. Los mestizos son el resultado de este proceso; según el censo de 1921 el 59.3 % de la población se autoidentificó como mestiza (Informe Cepal de Juan Cristóbal Rubio Badán https://repositorio.cepal.org/bitstream/handle/11362/36858/1/S1420252_es.pdf).

Durante la época colonial, tres epidemias de viruela, sarampión y tifus disminuyeron drásticamente la población indígena en el siglo XVI (Escalante Gonzalbo *et al.*, 2004). En este período, una importante cantidad de comunidades y grupos étnicos desaparecieron por completo, aunque la densidad de población nativa se recuperó algún tiempo después. Un evento que contribuyó a la mezcla fue el desarrollo industrial promovido por el descubrimiento de minas de plata en el norte de México del siglo XVI al siglo XVIII. Producto de este proceso, aparecieron

asentamientos extranjeros en todo el territorio mexicano y, en consecuencia, la población mestiza ha aumentado significativamente en los últimos 500 años. Además, la inmigración reciente de los centros rurales a los centros urbanos ha aumentado las diferencias regionales de mezcla, principalmente el aumento de ascendencia nativo-americana en centros urbanos.

En México, los datos arrojados por la Encuesta Nacional de Dinámica Demográfica (CONAPO, 2009) realizada por el Consejo Nacional de Población (CONAPO) estimaron que hay 14.3 millones de habitantes indígenas en el país, representando aproximadamente el 11 % de la población total. Sin embargo esta estimación tiene un aspecto a considerar, ya que el parámetro utilizado para determinar la pertenencia a un grupo indígena fue estimado mediante dos preguntas: *"¿Habla alguna lengua indígena o dialecto?"* y *"¿Qué lengua indígena o dialecto habla?"* (Rubio Badán, 2014). Si bien el lenguaje es un factor determinante del sentido de pertenencia a un grupo (Rodríguez Sala-Gómezgil, 1983) el mismo se limita a la identidad cultural, en cambio, cuando el foco se centra en la biología con un abordaje evolutivo es necesario utilizar parámetros que recaben aquellas características biológicas que las poblaciones transmiten a través de las sucesivas generaciones.

Los estudios de diversidad biológica en México, en particular los que consideran diversidad genética, son muchos y variados, pero casi todos concluyen que las poblaciones nativas presentan niveles bajos de diversidad genética, más bajos aún que otras poblaciones continentales pero muy divergente entre sub-poblaciones (Moreno-Estrada *et al.*, 2014; Wang *et al.*, 2007).

Los primeros estudios estuvieron limitados en el número de *loci*, en el número de subpoblaciones y en los tamaños muestrales (Lisker *et al.*, 1996), otros analizaron únicamente ADN mitocondrial (Gorostiza *et al.*, 2012) o cromosoma Y (Sandoval *et al.*, 2012); posteriormente se amplió este espectro y comenzaron a publicarse estudios con mayor cobertura genómica y tamaños poblacionales más grandes, como es el caso del estudio publicado por Moreno-Estrada *et al.* (2014) que muestra la diversidad étnica de México. Este estudio analiza 1 millón de SNPs

en 20 subpoblaciones de pueblos originarios (511 individuos) y 11 poblaciones cosmopolitas de México; entre las subpoblaciones de pueblos originarios tenemos representantes de todo el territorio mexicano (Norte de México: seris, tarahumaras, tepehuanos y huicholes; Centro-Oeste de México: purepechas y nahuas de Jalisco; Centro-Este de México: totonacas, nahuas de Puebla y de Trios; Sur de México: nahuas de Guerrero, mazatecos, triquis, zapotecas; y Sur-Este de México: mayas, tzotziles, tojolabales y lacandones). El estudio confirma el alto grado de diferenciación entre poblaciones y concluye sobre el efecto de esta estructura en los estudios biomédicos y su potencial efecto confusor en los resultados (Moreno-Estrada *et al.*, 2014).

Este último estudio (Figura 1.2, tomada de Moreno-Estrada *et al.*, 2014) muestra la variabilidad de ancestrías entre poblaciones cosmopolitas de México donde se puede observar un gradiente norte-sur y este-oeste de las proporciones de los componentes parentales. Las poblaciones del sur de la costa Pacífico presentan los mayores niveles de ancestría nativa y a medida que nos movemos hacia el norte la misma disminuye y aumenta la ancestría europea. En el caso de la ancestría africana, en todas la poblaciones estudiadas es la minoritaria. Todas estas consideraciones a la vez que muestran la estructuración poblacional compleja, nos obligan a considerar esta situación en estudios biomédicos donde se buscan los polimorfismos responsables o asociados a determinados rasgos de interés (Moreno-Estrada *et al.*, 2014).

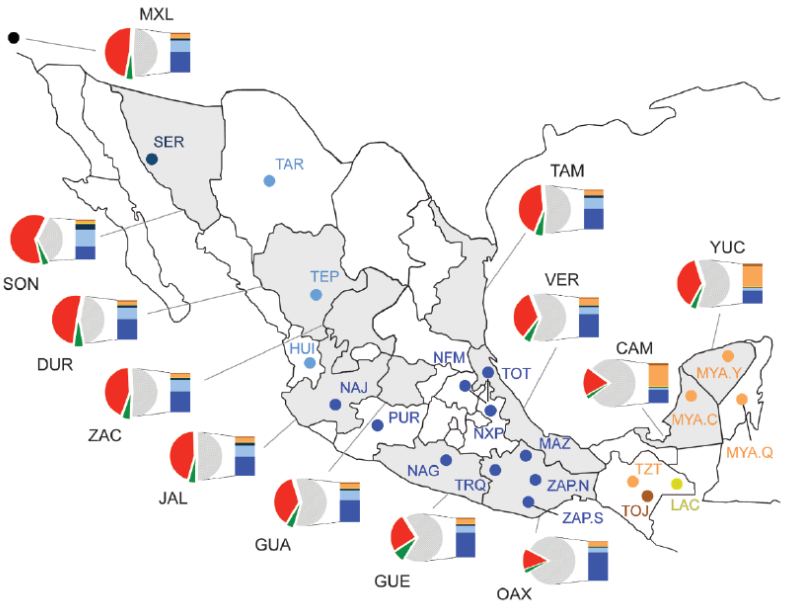

Figura 1.2: Estructura de la población mexicana. Los gráficos sectoriales muestran las proporciones promedio, por estado, de muestras cosmopolitas estimadas usando k=3 (en rojo europea, africana en verde y en gris nativo-americana). Las muestras corresponden a los estados de Yucatán (YAC), Campeche (CAM), Oaxaca (OAX), Veracruz (VER), Guerrero (GUE), Tamaulipas (TAM), Guanajuato (GUA), Jalisco (JAL), Zacatecas (ZAC), Durango (DUR), Sonora (SON) y una población de Mexicanos residentes en Los Ángeles (MXL) (Tomada de Moreno-Estrada *et al.*, 2014).

1.3.2. Poblaciones mestizadas y su interés en estudios de cáncer

El aporte de los GWAS es importante ya mapean regiones genómicas que podrían estar afectando al riesgo a padecer la enfermedad; sin embargo, los estudios de GWAS han explicado solo una parte de la heredabilidad de las enfermedades estudiadas (Eichler *et al.*, 2010; Manolio *et al.*, 2009), y no es asunto del tamaño muestral de los estudios sino que, el riesgo relativo asociado a la mayoría de los SNPs es modesto y muchas variantes podrían localizarse fuera de los límites del gen, así como también puede suceder que estos SNPs están en muy baja frecuencia en la población estudiada. En la Figura 1.3 se muestra la relación entre la frecuencia del alelo y el grado de penetrancia del polimorfismo sobre la enfermedad estudiada.

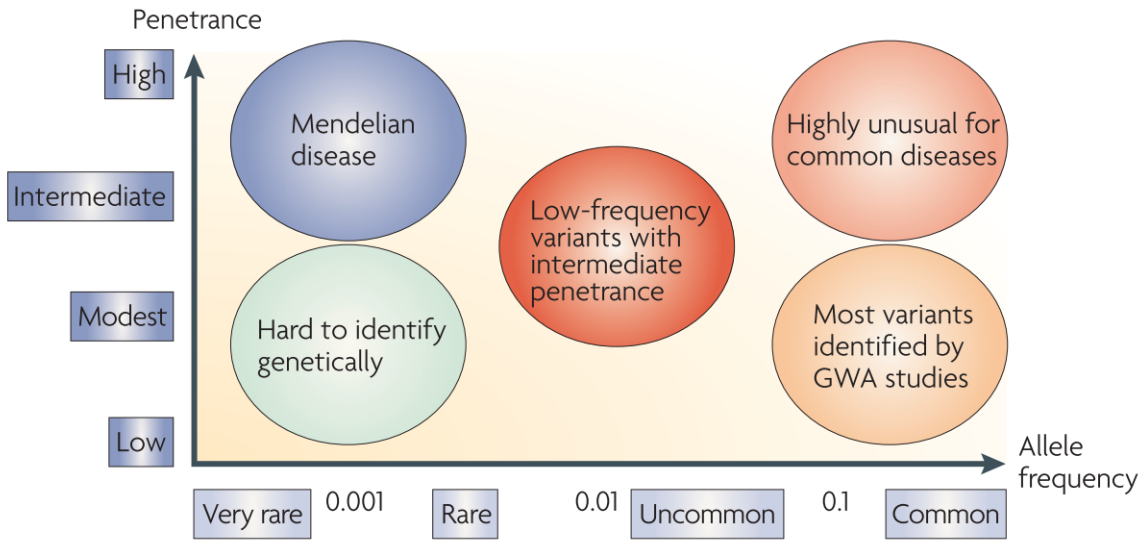

Figura 1.3: Frecuencia alélica en función del grado de penetrancia de la variante estudiada. Tomada de McCarthy *et al.* (2008).

Se ha observado que frente a polimorfismos cuya alelo minoritario tiene una frecuencia menor a 5 % (MAF, por sus siglas en inglés), el odd ratio (OR) es modesto o pequeño, pero aunque modestos, al combinarse pueden dar cuenta de un efecto considerable. Las poblaciones mestizadas de Latinoamérica (LA) ofrecen un escenario alternativo al trabajo con poblaciones no mestizadas, ya que se caracterizan por haberse formado recientemente (poco más de cinco siglos de mestizaje). Por esta razón, es un escenario idóneo para identificar polimorfismos que no son detectables en estudios con poblaciones no mestizadas, ya que la mezcla genera nuevas combinaciones genotípicas no presentes en las poblaciones parentales. Esta idoneidad se debe mayoritariamente a dos razones, la ocurrencia de alelos no polimórficos o "raros" (baja frecuencia) en las poblaciones parentales y a los bloques haplotípicos en alto desequilibrio de ligamiento (DL) generados por el mestizaje reciente. Respecto a las frecuencias alélicas, los polimorfismos tienen frecuencias variadas en las diferentes poblaciones y aquéllos más recientemente generados presentan mayores niveles de diferenciación entre las poblaciones (Gravel *et al.*, 2011; Mathie-

son & McVean, 2012); en este sentido, las poblaciones mestizadas, por su historia demográfica particular, permiten detectar polimorfismos que no son observados en otras poblaciones por estar en una frecuencia baja o fijos. Son estos polimorfismos, los que permiten complementar los estudios tradicionales de GWAS e identificar SNPs de baja frecuencia asociados al rasgos de interés.

Por su parte, las estructuras en bloque son una potente herramienta para los estudios genético-epidemiológicos, no solo por su importancia en determinar los procesos históricos que atravesaron estas poblaciones sino por la capacidad de dilucidar el escenario genético subyacente en enfermedades complejas, entre las cuales se encuentra el cáncer. Los bloques generados por DL aumentan la probabilidad de detectar el alelo causante ya que no hace falta genotipar directamente este alelo sino que basta con genotipar otro polimorfismo del mismo bloque haplotípico en alto DL con éste. Esta situación se ve favorecida al trabajar con genomas de individuos latinoamericanos, ya que en ellos se observan bloques de ligamiento más extensos que en sus poblaciones parentales (Chakraborty & Weiss, 1988; Freedman *et al.*, 2004; McKeigue, 1998; Patterson *et al.*, 2004; Pfaff *et al.*, 2001; Service *et al.*, 2006; Smith *et al.*, 2004). Este patrón se describe en nuestro trabajo publicado recientemente (ver capítulo 3), el cual muestra que los individuos de México tienen menor cantidad de bloques de ligamiento pero en promedio más largos que otras poblaciones mundiales (Colistro *et al.*, 2020).

También podemos destacar otra ventaja de trabajar con poblaciones mestizadas, como ser la presencia de ciertos alelos que pueden ser más fácilmente detectables por conferir mayor riesgo relativo en LA que en Europa, debido a interacciones gen-gen o gen-ambiente específicas de la población.

Estas particularidades de los genomas de individuos mestizados han sido estudiadas por varios autores y no solo concluyen sobre las ventajas de realizar estudios en estas poblaciones sino que ponen en evidencia la importancia de incorporarlas sistemáticamente en estudios de asociación para poder generar un conocimiento más amplio en la materia (Bryc *et al.*, 2010;

Wang *et al.*, 2008). Un estudio llevado adelante en poblaciones de LA (México, Colombia, Perú y Puerto Rico) identifica un set de SNPs localizados en genes de vías metabólicas del sistema inmune, con frecuencias anómalas respecto a las frecuencias esperadas según la composición ancestral de cada población, lo cual según los autores, es una evidencia de fuerzas de selección similares durante la evolución de estas poblaciones con efectos sobre fenotipos relacionados a salud y enfermedad (Norris *et al.*, 2018).

Puntualmente, poblaciones de mestizos mexicanos pertenecientes a seis estados (Sonora, Zacatecas, Guanajuato, Guerrero, Veracruz y Yucatán) fueron estudiadas y se concluyó que la estructura haplotípica de población mexicana, por sus características de mestizaje, es idónea para reducir la cantidad de SNPs que se requieren genotipar sin perder potencial de detección de asociación en estudios biomédicos y de diversidad (Silva-Zolezzi *et al.*, 2009).

1.4. Hipótesis y objetivos

1.4.1. Hipótesis

Sucesivos GWAS no han podido explicar un gran porcentaje de la variabilidad y etiología de las enfermedades estudiadas. Se plantea como hipótesis de trabajo que las variantes asociadas a CRC determinadas en poblaciones no mestizadas son en parte diferentes a las que se asocian a esta enfermedad en poblaciones mestizadas.

1.4.2. Objetivo General

El presente estudio propone aportar a dilucidar el impacto que tiene la historia demográfica de mestizaje sobre la variabilidad genómica de la población de México y el efecto en la asociación de regiones genómicas a una enfermedad compleja. Para esto se toma como ejemplo una muestra de la población mexicana y una enfermedad compleja, el cáncer colorrectal.

1.4.3. Objetivos Específicos

1. Determinar las proporciones de ancestría genética de una muestra de individuos de México y analizar el impacto de sobre los patrones de variabilidad genómica en regiones previamente asociadas a CRC.

2. Examinar las estimaciones de parentesco en muestras de casos y controles de CRC de México utilizando diferentes abordajes metodológicos y analizar las respectivas ventajas y desventajas al trabajar con poblaciones mestizadas.

3. Analizar si los SNPs que presentan asociación a CRC en Europa también lo hacen en la población de México y determinar si se detectan nuevos SNPs no antes descritos en

poblaciones parentales.

4. Determinar si los genes cercanos a los nuevos SNPs asociados muestran asociación al considerar el efecto combinado de todos los SNPs del gen, tanto los genotipados directamente como los imputados.

Capítulo 2

Análisis de ancestría en genes candidatos

2.1. Mestizaje diferencial de poblaciones Latinoamericanas y su impacto en el estudio del cáncer colorrectal.

Este capítulo consta de un artículo publicado en la revista *Genetic and Molecular Biology* (Colistro *et al.* 2020, ver Anexo I). En él se aborda el objetivo específico 1, el cual plantea determinar las proporciones de ancestría genética de una muestra de individuos de México y analizar patrones de variabilidad genómica en regiones previamente asociadas a CRC.

Se analizaron las proporciones globales de ancestría de los individuos mexicanos, para ellos se consideraron genotipos de individuos con ancestrías similares a las poblaciones parentales de la población Mexicana, disponibles públicamente en la plataforma *1000G*.

Luego, identificamos 10 regiones genómicas descritas como los principales responsables del CRC y analizamos patrones de variabilidad genómica entre muestras genotípicas de poblaciones mundiales y genotipos de muestras de México colectadas en el marco del proyecto CHIBCHA. Estos análisis los hicimos a través de estimaciones de ancestría local de las 10 regiones, de

heterocigosidad media global, heterocigosidad media discriminada por la categoría del SNP (consecuencia del SNP para le transcripto) y finalmente por cantidad de bloques de ligamiento en esas regiones y el largo de los mismos.

Los resultados de ancestría global de la muestra fueron consistentes con estudios previos, los cuales son choerentes con la historia demográfica de la población de México. Del análisis de ancestría por región genómica se observa una gran heterogeneidad de escenarios, en algunas regiones la ancestría predominante fue la europea (16q22.1) y en otras (*MLH1* y *MSH6*) la asiática (como representante de la población Nativo Americana). Esta heterogeneidad nos inhibe de poder proponer generalizaciones y evidencia la complejidad de los análisis con poblaciones mestizadas. Lo mismo ocurre al observar los resultados de los análisis de heterocigosidad. Por último, del estudio de los haplobloques se desprende que el tamaño de los mismos apoya la idea que las poblaciones mestizadas tienen mayor valores de desequilibrio de ligamiento.

Todos estos resultados muestran que cada región tiene un comportamiento diferente y que las generalidades deben ser cautelosas ya que de no tenerse en cuenta las particularidades de cada región, podría conducir a resultados espúreos, producidos por cálculos sesgados dada la no consideración de la naturaleza mestizada de las muestras.

En este trabajo tuve un rol principal en todas la etapas que llevaron a la publicación de este artículo. Estas etapas comprenden el proceso de curado y puesta a punto de los datos genómicos de las muestras mexicanas, la minería de datos de las plataformas mencionadas, los análisis *in sillico*, el procesamiento bioinformático de las bases de datos, los análisis estadísticos, la visualización de los resultados y la redacción del manuscrito. Este trabajo fue realizado en conjunto con los co-autores, en particular con la Dra. Mónica Sans cuya participación fue crucial en todas las etapas mencionadas, destacándose la conceptualización del trabajo.

Differential admixture in Latin American populations and its impact on the study of colorectal cancer.

Colistro V.[1], Mut P.[2], Hidalgo P.C.[3], Carracedo A.[4,5], Quintela I.[4], Rojas-Martínez A.[6], Sans M. [*,2].

[1] Universidad de la República, Facultad de Medicina, Departamento de Métodos Cuantitativos, Montevideo,Uruguay.
[2] Universidad de la República, Facultad de Humanidades y Ciencias de la Educación, Departamento de Antropología Biológica, Montevideo, Uruguay
[3] Universidad de la República, Centro Universitario de Tacuarembó, Polo de Desarrollo Universitario Diversidad Genética Humana, Tacuarembó, Uruguay
[4] Universidad de Santiago de Compostela, Centro Nacional de Genotipado (CEGEN), Spain
[5] Universidade de Santiago de Compostela, CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud Carlos III, Grupo de Medicina Xenómica, Santiago de Compostela, Spain.
[6] Escuela de Medicina y Ciencias de la Salud, Tecnológico de Monterrey, Monterrey, México
*Corresponding authors.

Abstract

Genome-wide association studies focused on searching genes responsible for several diseases. Admixture mapping studies proposed a more efficient alternative capable of detecting polymorphisms contributing with a small effect on the disease risk. This method focuses on the higher values of linkage disequilibrium in admixed populations. To test this, we analyzed 10 genomic regions previously defined as related with colorectal cancer among 8 populations and studied the variation pattern of haplotypic structures and heterozygosity values on seven categories of SNPs. Both analyses showed differences among chromosomal regions and studied populations. Admixed Latin-American samples generally show intermediate values. Heterozygosity of the SNPs grouped in categories varies more in each gene than in each population. African related populations have more blocks per chromosomal region, coherently with their antiquity. In sum, some similarities were found among Latin American populations, but each chromosomal region showed a particular behaviour, despite the fact the study refers to genes and

regions related with one particular complex disease. This study strongly suggests the necessity of developing statistical methods to deal with di or tri-hybrid populations, as well as to carefully analyze the different historic and demographic scenarios, and the different characteristics of particular chromosomal regions and evolutionary forces.

Introduction

One of the greatest challenges in genetic epidemiology is the development and application of methodological strategies allowing identification of genetic risk loci in order to achieve a more thorough understanding of the genetic basis of complex diseases, as they are the result of interactions between multiple genetic and/or environmental factors, each with modest effects. It is likely that different combinations produce the same clinical symptoms (Botstein and Risch 2003). Also, many complex diseases are genetically related, sharing common genetic risk variants (Teng *et al.* 2016). Moreover, interconnections among all genes expressed in disease-relevant cells and the core disease-related genes (omnigenic model) (Boyle *et al.* 2017).

Linkage and association studies are the two main approaches applied to identify the genetic basis of these types of diseases (Patel *et al.* 2003; Morton 2003; Khoury *et al.* 2010). Linkage studies are more efficient in detecting genes with large effects, like single-gene based disorders, but they lack the statistical power to detect variants with modest effects. On the other hand, genome-wide association studies (GWAS) have a statistical advantage as they provide greater power for detecting common variants with modest risk (Risch and Merikangas 1996). However, these studies have been criticized, as they rely on an extremely high number of markers in order to be carried out (more than 100.000), a large quantity of samples, as well as adequate technological resources to process the enormous amount of data, becoming impractical and very expensive (Cantor *et al.* 2010; Qin and Zhu 2012).

Admixture mapping studies (AMSs) constitute an alternative approach. This methodology was first proposed by Rife (1953), but its implementation has been technically possible only in the last decades (McKeigue 2005). AMS is based on the gene flow processes between continental

populations occurring in the last centuries, producing particular chromosome configurations in the resulting admixed populations, showing a mosaic of ancestry segments (Darvasi and Shifman 2005). When a disease has substantial despair prevalence among parental populations, the risk allele locus will show an over-representation ancestry of the high risk population in the admixed population. The use of ancestry informative markers (AIMs) allows the identification of the population source of the studied chromosomal segments (Tian *et al.* 2008; Winkler *et al.* 2010, among others). The effect of rare variants in recently admixed populations can be much greater compared with its ancestral populations, as has been shown by Moltke and Albrechtsen (2014). Moreover, the effects of noncausal genetic variants depend on its correlation with causal variants, and these last may vary depending on the ancestral populations and the patterns of linkage disequilibrium (Skotte *et al.* 2019).

The process of admixture in the Americas can be seen as a natural experiment for genetic epidemiology and anthropology, in which polymorphic marker loci are used to infer a genetic basis for traits of interest (Chakraborty and Weiss 1988). Nowadays it is possible to establish a maximum of approximately 21 generations of admixing, depending on the region. Nowadays, cosmopolitan Latin American populations have Native contributions from around 1 % to more than 50 %, and African contributions from 2 to 40 %, while on the other side, it is rare to find Native groups without any admixture (Sans 2000).

The grade of contribution of each parental population will be reflected not only in the amountof chromosomes from each ancestral origin, but in the quantity of blocks from these origins inside chromosomes and depend on the admixture process (Pfaff *et al.* 2001). We assume that the antiquity of populations is directly related to the heterozygosity and the size of chromosomal blocks; consequently, we expect smaller blocks in more ancient populations. Moreover, heterozygosity can be related to the time (generations) after a process of admixture, assuming that non-admixed populations are more homogeneous. We recognize that it is an oversimplification because it ignores the microevolutionary changes in the admixed population,

as genetic drift, selection and gene flow.

The major aim of our study was to understand the process that generates complex chromosome patterns in admixed populations and to improve the implementation of AMSs inLatin American populations. Particularly, this study was focused on analyzing genes and chromosomal regions previously related to colorectal cancer (CRC) in admixed populations, because past studies were mainly based on populations of European descent. CRC is common in both sexes and has no major avoidable risk factor. By determining the ancestral proportions, as well as the heterozygosity and size of fragments in five admixed American populations and several populations from Europe, Africa and Asia in associated regions, we intend to help in the understanding of genetic CRC causes.

Materials and methods

Samples

We used data available in 1000 Genomes Project for 8 populations and an unpublished set of genetically admixed Mexican samples. Regarding the 1000 Genome Project samples, five areadmixed populations from the Americas, and the others were selected to represent part of their parental populations. The admixed populations were: Afro-Americans from the United States (ASW, N=83), Colombians (CLM, N=60), Puerto Ricans (PUR, N=55), Peruvians (PEL, N=85) and Mexicans from Los Angeles, CA (MXL, N=76). The samples from Africa, Europe and Asia were selected due to their relationship to the migrations toward America, being the last ones considered in substitution of Native Americans. We are aware of differences between Asian and Native American populations, but we choose this alternative due to the scarcity of data referred to the SNPs and regions considered for such populations. Therefore, we analyzed Yorubas and Luhya to represent African populations (denominated Africans, AF, in this paper, N=176), Iberians, Tuscans, and Utah residents with northern and western European ancestry for European populations (denominated EU, N=174), and Chinese from Beijing, Southern Han Chinese and Japanese from Tokyo to represent Asians (denominated AS, N=98).

We are particularly interested in another Mexican sample (MEX, N=831) because it is formed by healthy controls of a GWAS study of CRC (CHIBCHA, study of hereditary cancer in Europe and Latin America). The individuals were recruited in different blood banks, three in México City (Centro Médico Nacional Siglo XXI of the Mexican Social Security Institute -IMSS-), three in Monterrey (UMAE 25, IMSS and the University Hospital of the Universidad Autónoma de Nuevo León) and three in Torreón (UFM 16 IMSS, the UMAE 71 IMSS, and the University Hospital of Torreon), from 2010 to 2012. All subjects gave informed consent for inclusion before they participated in the study. The protocol was approved by the ethics committees of each participating institution (Ethics Committee of the University Hospital, Universidad Autónoma de Nuevo León code BI10-003 and the National Commission of Scientific Research of the Mexican Social Security Institute code R-2012-785-032), the Federal Commission for Protection against Health Risks (COFEPRIS), code CMN2012-001, and the Ethics Committee of CHIBCHA project number: 223 678. Samples were genotyped using two complementary arrays: Axiom Genome-Wide LAT 1 (Latino) Array and a Custom-designed Array, both from Affymetrix Axiom Genotyping Solutions. The former was designed to maximize coverage of common and rare disease-associated alleles in Latin American populations that have genetic contributions from European, Native American and African ancestries. The latter was specifically designed for this study, being the SNPs selection based on regions previously detected as associated with CRC. SNP callingin both arrays was done following Affymetrix best practice workflow, which includes the Genotyping Console Software in combination with SNPolisher. A total of 1,169,944 SNPs (387,948 from the Custom Array and 781,996 from the Latino Array) was obtained. These samples were included because its large number of individuals and the high coverage of SNPs in the considered regions, represent an opportunity to compare the performance of another admixed population.

Genotypes of Native American (NAM) samples were used in order to estimate the global individual ancestry. These genotypes included individuals from five ethnic groups: Zapotecs from Oaxaca, Mexico (N=21), Tepehuans from Durango in Northern Mexico (N=23), Nahuas

from Central Mexico (N=14), Mayas from Campeche, Mexico (N=25), Quechuas from Cerro de Pasco, Perú (N=24) and Aymaras from La Paz, Bolivia (N=25). We consider apanel of AIMs developed and optimized for the study of Latin American populations by the LACE Consortium (for detailed information about the panel and the populations refer to Galanter *et al.* 2012). This panel was composed of 446 AIMs but the ancestry analysis performed in the present study was limited to the 275 SNPs shared with the Mexicans, the 1000G populations and the Native American samples.

Genomic regions studied

We selected 10 autosomal regions, with an average size of 680.9 Kbp spanning a total of 6.8 MB (Table 1). These regions were previously described to show association to CRC (Kinzler *et al.* 1991; Aaltonen *et al.* 2007), 7 of them are genes: *APC*, *BRAF*, *MSH2*, *MSH6*, *MLH1*, *MUTYH* and *PMS2*, and three are loci described by Carvajal-Carmona *et al.* (2011) also associated with CRC: 8q23.3, 16q22.1 and 19q13.11.

For the seven gene regions, SNPs within the gene limits were retrieved, and in the three otherregions 1 MB upstream and downstream SNPs were considered. The number of available SNPs in each region is listed in Table 1.

Admixture Analysis

In order to understand the structure of the MEX sample, we performed a global individual admixture analysis using the AIMs panel described above. Estimation of individual admixture fractions were calculated with ADMIXTURE software version 1.3.1 (Alexander etal. 2009), which considers a likelihood model. To choose the correct value of k we computed the cross-validation error over k, from 2 to 6. We found that k=3 yielded the lowest cross-validation error (k_3=0.538) compared to other k values (k_2=0.63968, k_4=0.54016, k_5=0.54226 and k_6=0.542).

Complementary, we also analyzed the mean population admixture in each of the 10 regions for the admixed populations. In this case we were not able to use the Native American samples

Table 1. Genomic regions considered in the analysis, 7 genes and 3 locations (± 1 MB). The table shows chromosome, base pair start and end, gene name, cytoband and number of SNPs of each studied location.

Chromosome	Cytoband	Gene Start (bp)	Gene End (bp)	Gene name	SNPs
1	p34.1	45794835	45806142	MUTYH	706
2	p21	47630108	47789450	MSH2	1058
2	p16.3	48010221	48034092	MSH6	1011
3	p22.2	37034823	37107380	MLH1	997
5	q22.2	112043195	112181936	APC	1140
7	q34	140424943	140624564	BRAF	1138
7	p22.1	6012870	6048756	PMS2	682
8	q23.3	116631278	118626279	—	864
16	q22.1	67824395	69816284	—	964
19	q13.11	32534093	34530086	—	1025
Total					9585

due to their limited number of SNPs yielding at these 10 regions. As explained above, we used the Asian samples instead. A total of 5283 SNPs were used for this analysis.

Analysis of Genetic Variation

The genetic variation analysis was performed only on the 7 genomic regions corresponding togenes. To compare the variation in the studied regions among the 9 populations, we considered two measures using PLINK version 1.9 (Purcell *et al.* 2007; Chang *et al.* 2015): heterozygosity and haplotypic structures among regions and populations.

For the heterozygosity determination, the mean values of heterozygosity were analyzed for each gene by population and the mean values of SNPs were classified in seven categories. The SNPs classification categories are related to their position and consequence to transcript and were obtained using Biomart (Haider *et al.* 2009): intronic, non-synonymous coding, synonymous coding, 5' UTR, 3' UTR, stop gained and stop lost.

Inference of haplotype phase was determined with the Beagle software version 4 (Browning and Browning 2007). Gabriel *et al.* (2002) criteria was followed to define haploblocks. The allelic

association between pairs of SNPs was measured by the D' parameter (Lewontin 1964). The distribution of blocks length (in bp) among populations was compared. Linkage analysis and haploblock estimation were done using PLINK version 1.9 (Purcell *et al.* 2007).

Results

Admixture analysis

The AIM panel accurately discriminates parental populations,as can be seen in Figure 1a. The representation of the global individual ancestral fractions for the admixed populations is shown in Figure 1b and 1c. According to the estimations, the ASW population has 75,4 % of African ancestry, while the African proportions for the other admixed populations were lower: 12,1 % in CLM, 6,8 % in MXL, 4,3 % in PEL and 16 % in PUR. Peruvian samples (PEL) have the highest proportions of Native American ancestry (77,1 %) followed by the Mexican samples (MXL and MEX) (51,2 and 61,5 %, respectively). The European ancestry has its maximum in Puerto Rico (68,7 %) followed by the Colombian sample (61,7 %) (Table 2).

Table 2. Mean values of heterozygosity by population and by region. Highest and lowest values of each row were marked in order to facilitate visualization.

Gene	AF	ASW	AS	EUR	PUR	CLM	PEL	MEX	MXL
APC	0,063	0,066*	0,046	0,056	0,049	0,047	0,0378†	0,044	0,044
BRAF	0,079*	0,080	0,046	0,036	0,048	0,041	0,031	0,027†	0,029
MLH1	0,080	0,094*	0,016†	0,059	0,052	0,052	0,052	0,057	0,061
MSH2	0,063	0,065*	0,049†	0,052	0,054	0,061	0,052	0,052	0,054
MSH6	0,053	0,051	0,025†	0,082*	0,079	0,078	0,035	0,047	0,049
MUTYH	0,031	0,026	0,039*	0,024	0,032	0,032	0,020†	0,026	0,028
PMS2	0,094	0,101*	0,079	0,080	0,080	0,071	0,077	0,075	0,068†
Total	0,071	0,073*	0,045	0,052	0,054	0,051	0,042†	0,044	0,045

† lowest value in row; *highest value in row

When the admixture analysis was performed on those 10 regions considered in this study, the results show high variability (Figure 2). No clear pattern is detected among the different regions. In general terms there is a great concordance among populations than among genes

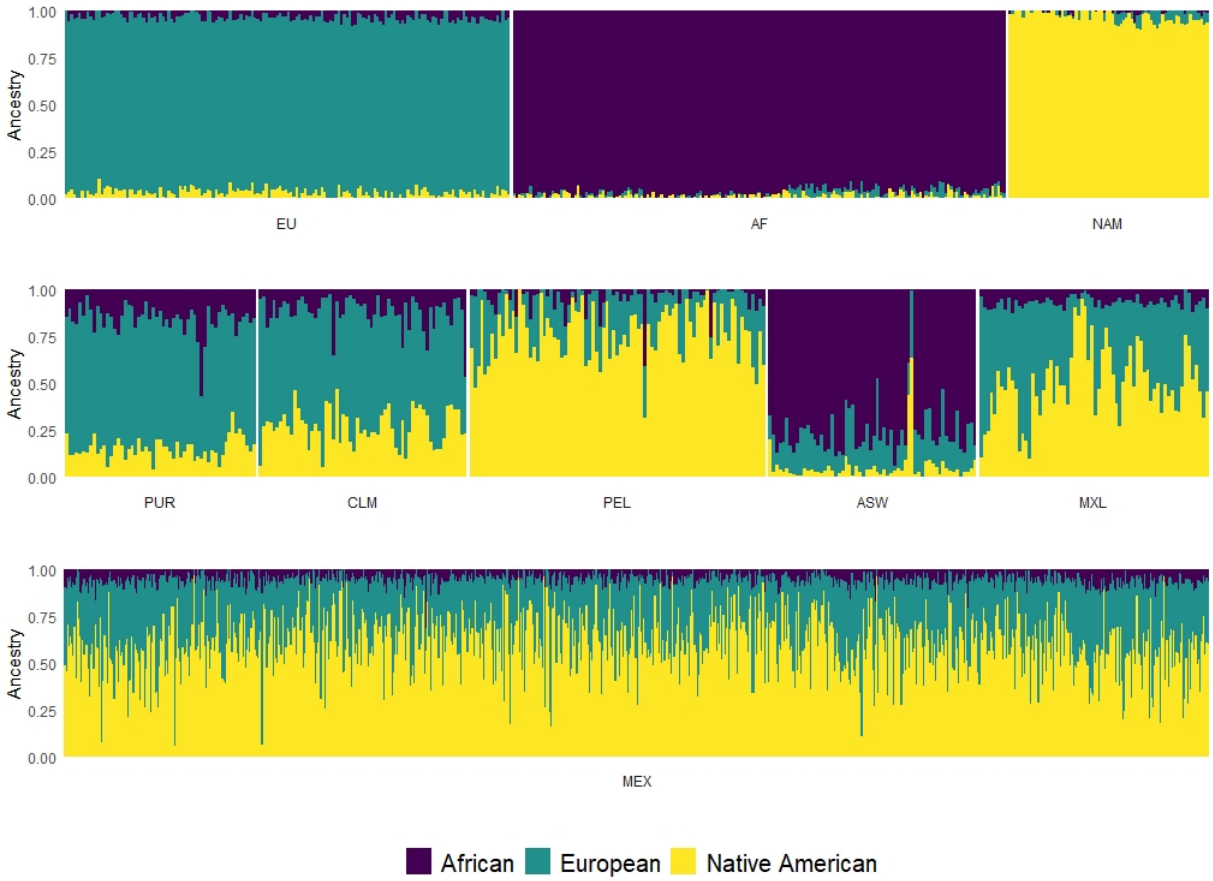

Figure 1. Global admixture analysis performed in ADMIXTURE, with k_3 representing the 3 ancestral components of the Admixed American populations. The barplots show each individual as a vertical line, and the ancestries are indicated by different colours (NAM= Native American ancestry, AFR= African ancestry and EUR= European ancestry). a) Parental populations, b) Admixed populations from 1000G and c) Mexican unpublished samples.

and regions. The greatest similarity is between both Mexican samples, while Peruvians seems to be the most dissimilar. While in *MSH6* and *MLH1* genes, a greater contribution of Asian ancestry was detected, and in 16q22.1 and *MUTYH* the European contribution is the highest.

Genetic Variation

The results of the analyses of the mean heterozygosity by gene are shown in Table 2 and the mean heterozygosity using the categories of SNPs mentioned above are shown in Figure 3.

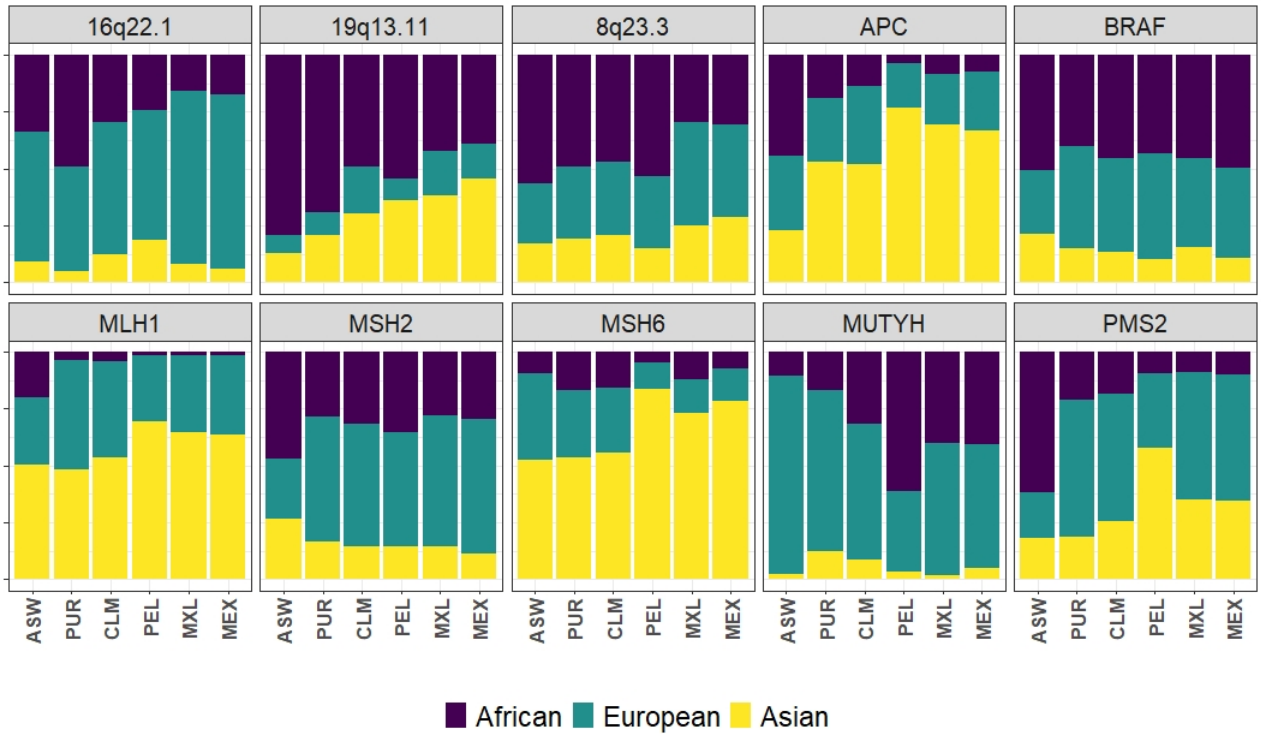

Figure 2. Admixture analysis by region performed with ADMIXTURE, with k_3 representing the 3 ancestral components of the Admixed American populations. The barplots show the mean ancestry of each population, and the ancestry proportions are indicated by different colors.

For two of these categories (stop gained, stop lost), no population showed heterozygosity in any region.

The greatest mean values of heterozygosity for most of the genes are found in the ASW, except for *BRAF*, MSH6 and *MUTYH* where the greatest values are in AF, EU and AS respectively. And the lowest values are found in AS for *MLH1*, *MSH2* and *MSH6*; in PEL for *APC* and *MUTYH*, in MEX for *BRAF* and in MXL for *PMS2* (Table 2).

When including the SNP category in the analysis, different genes show different situations: a) heterozygosity related to categories of SNPs vary in different regions; b) some chromosomal regions do not show heterozygosity in some categories of SNPs; c) heterozygosity varies when considering different populations but its behaviour is relatively coherent in the different

categories: Africans and Afro-descendants, European and Asian, and Latin American admixed ones.

For example, for the *APC* gene, relative differences in heterozygosity values associated to the categories of SNPs remain constant in all populations being 3'UTR the one with greatest values of heterozygosity followed by synonymous variants, except in the EU and CLM samples, where synonymous variants are greater (Figure 3).

Regarding the *BRAF* gene, the diversity among populations is clear. For this gene, the African related samples (AFR and ASW) have higher values of heterozygosity in intronic and synonymous categories, while 3'UTR regions are more homogeneous. Puerto Rico has an intermediate place between African related and other considered populations (Figure 3).

The *MSH2* locus differs from the others analyzed. The 3'UTR SNPs show none or very small heterozygosity in every population, except for the AF, ASW and PUR samples. As populations and MXL do not have heterozygosity in 3'UTR and synonymous mutations, while MEX shows very little heterozygosity in those regions. Intronic SNPs show the higher heterozygosity in every population (Figure 3).

It is important to note that the admixed Latin-American samples (PEL, CLM, MXL, MEX and PUR) show heterozygosity values for all genes that tend to be intermediate among the values of the parental samples (EU, AS and AF). Although, the ASW, also admixed, shows a pattern closer to the AF than to any other sample, in concordances with the high contributions of African genes (76 %); in some cases, also PUR approximates more to those samples.

The quantity of phased haplotype blocks per gene was analyzed for each population (Figure 4). The African populations (AF and ASW) have more blocks per region for most of the genes, while Asians (AS) have fewer, followed by MXL, probably because of the high NativeAmerican contribution of Native genes (62 %), and by Europeans. All populations have a similar curve for the 7 genes, with some exceptions: CLM shows a large amount of haplotypeblocks in *BRAF*,

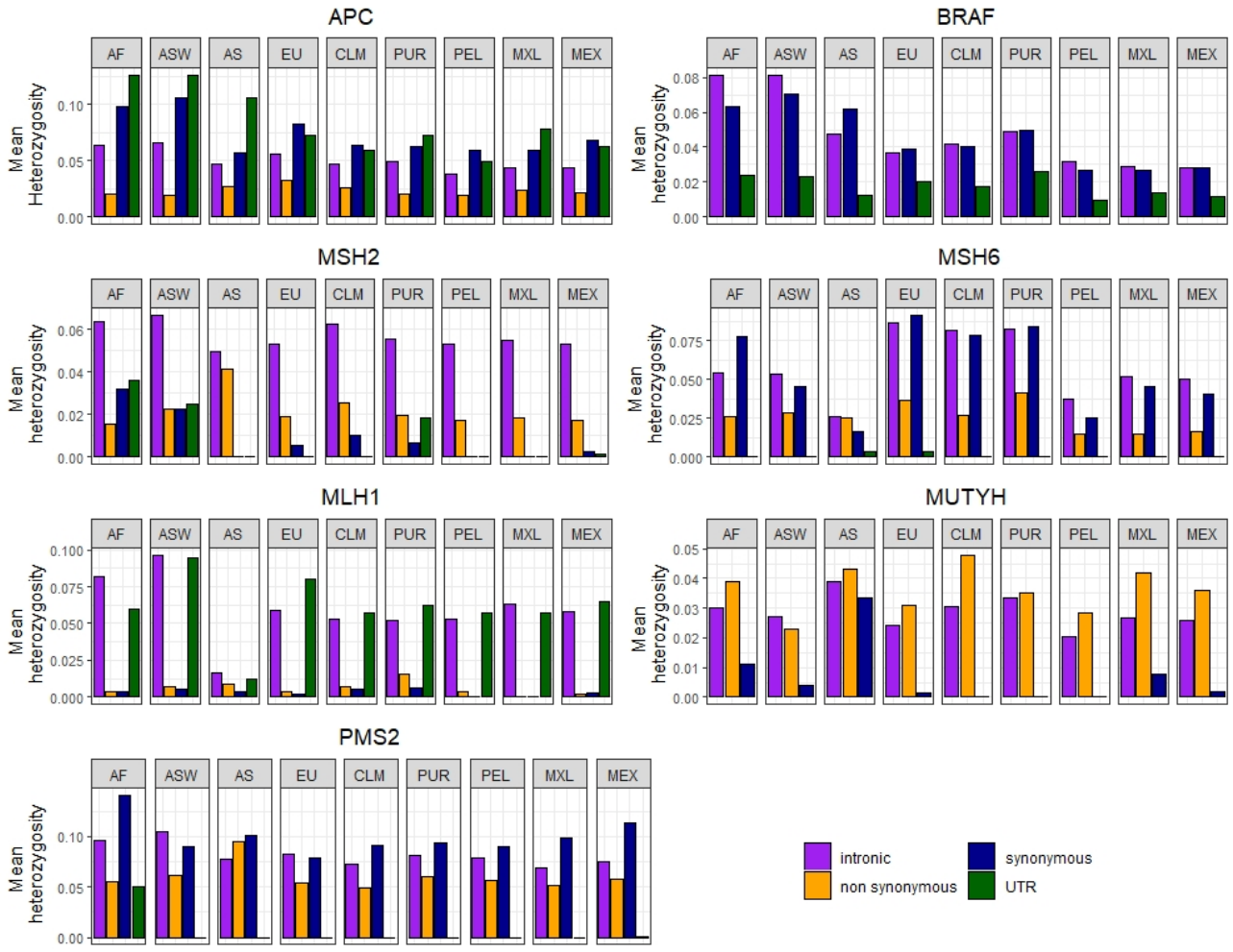

Figure 3. Mean heterozygosity values in four SNPs categories by gene and population. Each bar corresponds to a SNP category in a certain gene and population. SNPs categories are: intronic, non-synonymous, synonymous and UTR.

PUR that shows more blocks in *MSH6*, and MEX that shows more blocks at *MSH2* and *PMS2* genes and an unexpected behaviour related to the other Mexican sample (MXL).

The variability of the size of the blocks shows diversity among populations (Figure 5). It varied from <1 kb to over 190 kb, though most of the blocks were small (<5 kb). Markedly the African related populations (AF and ASW) have higher proportions of small blocks, and the admixed populations (CLM, MXL, MEX and PUR) are placed in an intermediate value between

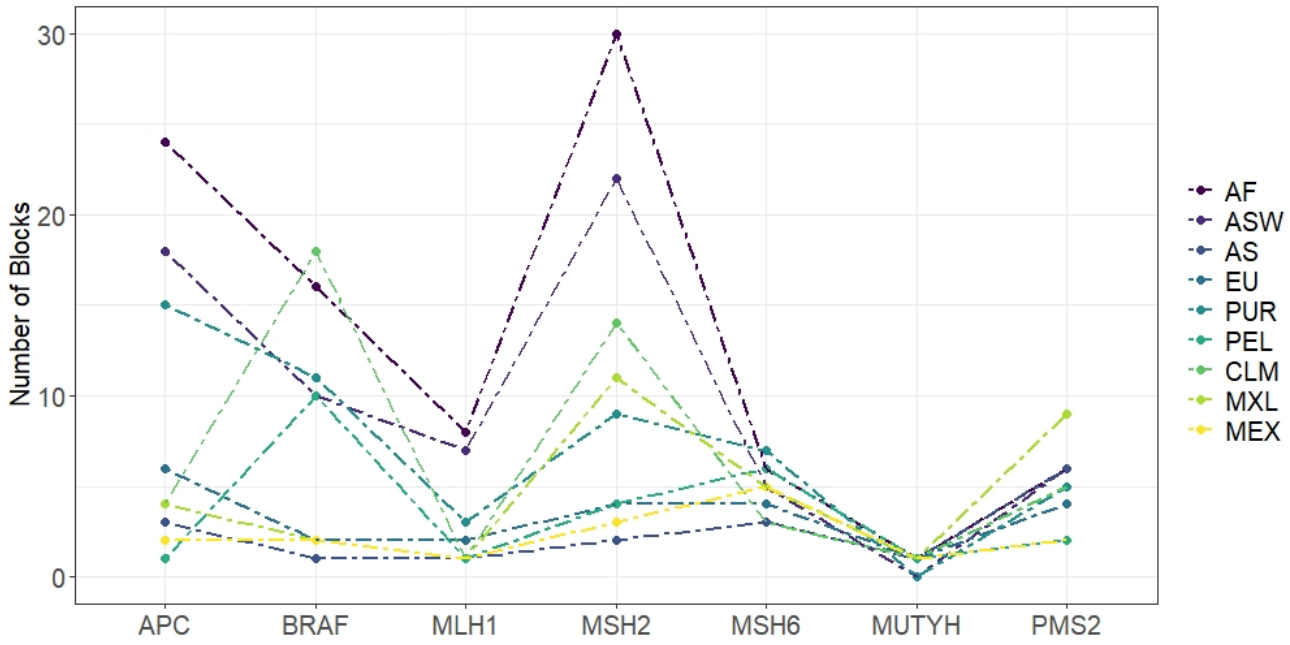

Figure 4. Number of phased haplotype blocks estimated for the 7 genes detailed per population.

the African-related and the other two parental populations (AS and EU). In the MEXsample, the smaller blocks are underrepresented in comparison with the rest of the admixed samples, while they show a greater number of longer blocks related to the rest of admixed populations.

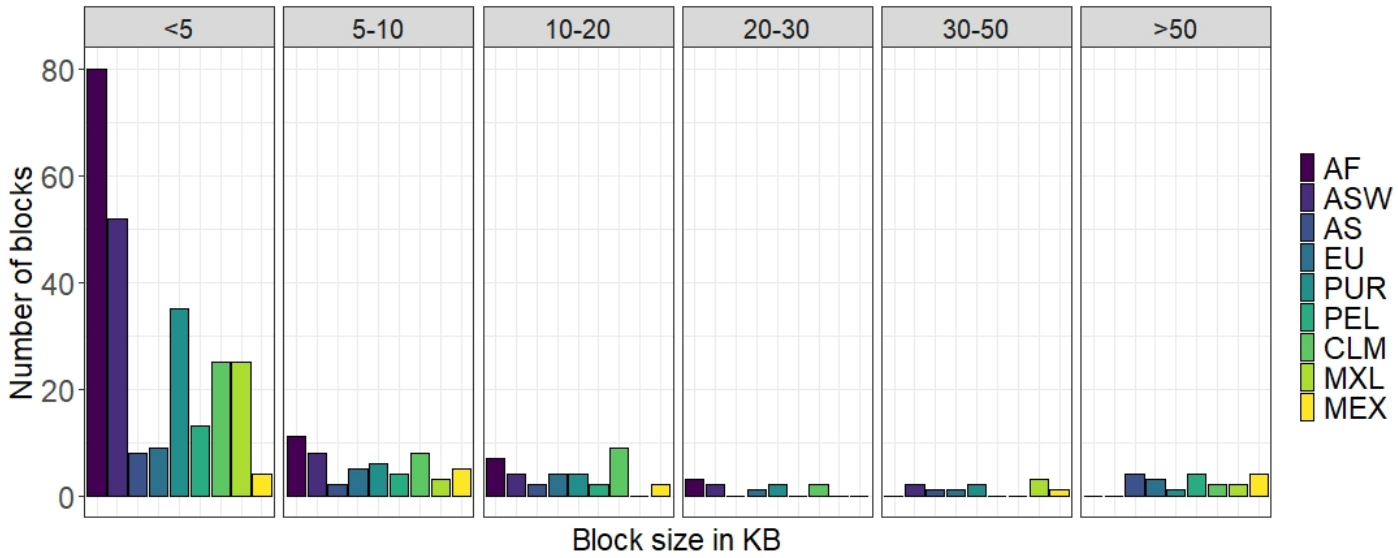

Figure 5. Block characteristics size (in kb) distribution of all haplotype blocks found in the analysis. Summary of haplotype diversity across all blocks.

Discussion

The results obtained using the selected AIMs supports the use of these markers for detecting admixture in Latin American populations, as demonstrated in several studies performed before (Mao *et al.* 2007; Tian *et al.* 2008; Halder *et al.* 2008; Silva *et al.* 2010; Galanter *et al.*2012; Manta *et al.* 2013). Moreover, we found that the expected proportions of ancestry are consistent with the historical and geographical affinities of the samples used, as well as other estimations (Norris *et al.* 2018). Peruvian and both Mexican samples showed the highest Native contribution, being 77,1 % in PUR, 51,2 % in MXL and 61,5 % in MEX; all three samples have the lowest African one (4-6 %). Different studies about population admixture in Mexico showed different contributions. In the central and northern regions, Native American contribution goes from 32 % to 69 % Native-American while African is usually less than 7 % (Martinez-Fierro *et al.* 2009; Salzano and Sans 2014). A comprehensive analysis by Rubi-Castellanos *et al.* (2009) in 10 Mexican regions shows somehow different results, presenting higher African contributions in some regions as Nueva Leon (18,5 %), Veracruz (17,2 %), and Jalisco and Campeche (15,9 %).

The ancestry analysis by region evidenced a different result in each one of the 10 regions. While in some genes the Asian contribution (as a proxy of Native American) predominates inall the admixed samples (*MLH1* and *MSH6*), in the 16q22.1 region the European contribution prevails. However, in most of the regions, the predominant ancestry is not the same for all samples. In *MSH2*, the European contribution is predominant except in the ASW in which theAfrican is the greatest. This exposes a different situation for each population and for each genomic region and outlines the importance of considering the local ancestry complementary to the global ancestry when performing association analysis in order to avoid spurious associations.

A similar conclusion can be drawn by taking into account the genetic variation analyses. The heterozygosity values showed very dissimilar ancestral contribution by population and by region. Only in one of the regions considered the highest and lowest mean values of heterozygosity were detected in one parental population (*MSH6*), being in most of the cases the highest mean value found in the African samples (all except MUTHY and *MSH6*). And finally, in four cases (*APC*, *BRAF*, MUTHY and *PMS2*), the lowest mean values were found in two admixed populations (PEL and MXL).

Both, in ancestry and in genetic variation analyses, the Native American contribution in Peruvian and Mexican samples is the highest, and consequently, it is possible to presuppose that the genetic variation patterns could be more closely related to Native Americans than in other Latin American populations. This is reflected in the *MSH2* gene heterozygosity values, as well as for haplotypic blocks of 5-10 kb, but not for the rest of the performed analyses. The non-expected values can be explained by different factors, like comparisons with Asian samples, instead of Native American samples. Moreover, some differences between the two Mexican samples were shown. The Mexican (MXL) sample was recruited in Los Angeles, California, and consequently, it can better be compared with Mexican Americans.

The MEX corresponds to the capital city, composed of subjects from the centre of the country, and Monterrey and Torreon, represented by subjects of northern parts of the country.

There is also a difference of 10 % of Native contribution, being greater in MEX than in MXL. Another crucial difference is the size of the samples (76 versus 831, respectively). This fact isnot minor, as bigger samples may uncover heterogeneities due to subestructuration. Then, variation in different parameters can be explained because of that, as the apparent presence ofvariation in heterozygosity at 3' and synonymous not found in MXL but in MEX in *MSH2* gene (Figure 2), or having more longer blocks shown in MEX sample (Figure 4). Also, differences between Mexican samples can be related to the coverage of the DNA analysis, being low for in MXL and high for MEX. It has been demonstrated by Ros-Freixedes *et al.*(2018) that low coverage can generate bias towards the detection of SNPs, showing that concordance with 10X coverage was 90,5 % for genotypes and 95,2 % for alleles, while with high coverage those values increased to 99,7 and 99,9 %, respectively.

The size of blocks supports that admixed populations have higher values of linkage disequilibrium that lead to a specific pattern of haplotypic structures. For example, PUR showed the higher values of European ancestry but despite that, its heterozygosity values are close to EUR for *BRAF* and *MSH2*, but not for *APC* or for haplotypes, where PUR are more similar to the other admixed samples.

Besides African and African-derived populations showed smaller blocks than the other populations, it is necessary to note that all populations analyzed here show a broad range of small blocks indicating little recombination in the regions, most genes, studied. As Gabriel etal. (2002) have demonstrated, African and African-American populations have around half ofthe genome concentrated in blocks of 22 kb or larger. Here we showed an intermediate situation in the Latin American population, despite some differences depending on the degree of admixture (and the origin of the genetic contributions) and the chromosomal region analyzed.

Two facts can be highlighted: 1) several evolutionary forces- not only genetic flow- act on genetic variability; and 2) each region analyzed has special behaviour when genetic variation is analyzed, despite all genes and chromosomal regions analyzed.

Related to the first, our data suggest that the patterns of ancestry and variability appear in certain genomic regions and under certain circumstances, but not in others. Different microevolutive forces such as selection, genetic drift, and eventually recombination, conversion and hitchhiking are probably present (Maynard Smith and Haigh 1974). Moreover, the evolutionary processes act on genetic regions and genes, being selection (positive or negative) the most important, followed by others as mutations (Salzano 2005). Besides, genetic flow is related to different migrations in the history of the involved populations that generated differences in populations and subpopulations (Stumpf and Goldstein 2003; Choudhry et al. 2006). Consequently, a deeper study taking into account historical and demographic scenarios as well as genetic variability is required before trying tomake inferences.

Related to the second, the 10 analyzed regions were detected as associated with CRC in European populations (Kinzler *et al.* 1991; Aaltonen *et al.* 2007; Carvajal-Carmona *et al.* 2011). Interestingly, when these regions were considered in the MEX sample when analyzing CRC in controls and patients, none of these genes showed association with the disease; only the 16q22.1 region was detected as associated (unpublished data). We would like to emphasize that our results suggest that not only global ancestry analysis is important when studying the association of genomic regions to a complex disease in admixed populations, but also regional ancestry analysis is advisable to be performed in order to detect an imbalance of ancestral contribution between cases and controls. Otherwise, associations might be the resultof the mentioned imbalance rather than the possible implication of that region in the disease considered.

Several authors (among others, Tishkoff and Verrelli 2003a; Tishkoff and Verrelli 2003b; González Burchard *et al.* 2005; Coop *et al.* 2009) have pointed out the importance of evolutionary factors (such as admixture) to understand the genomic structure of populations. Our data support that each population history and each genomic region needs to be studied independently. Consequently, we emphasize the importance of a prospective analysis of ancestral characteristics of the populations to be studied, especially when dealing with the admixed Latin

American populations where the di or tri-parental admix model is the most suitable.

Finally, this study strongly suggests the necessity of developing statistical methods to deal with di or tri-hybrid populations. It is also necessary to carefully analyze the different historical and demographic scenarios of each particular population to avoid generalizations, since, considering Latin America as a whole, is more theoretical than real.

Acknowledgments

To the members of the CHIBCHA Consortium Ian Tomlinson (University of Birmingham, UK), Luis Carvajal-Carmona (University of California, Davis, USA), Chris Holmes (University of Oxford, UK), Sergi Castellvi-Bel (Hospital Clinic, Spain), Manuel Teixeira (Portuguese Institute of Oncology, Portugal Magdalena Echeverry (Universidad del Tolima, Colombia) and Rocío Ortíz López (Tecnológico de Monterrey, México), and to the collaborators who take the Mexican samples. To the technician in Santiago de Compostela who collaborated in the genomic analyses. We are especially grateful to the people of Mexicowho participated in the study. This research was financed by the Seventh Framework Programme (FP7) of the European Commission, project number 223 678, "Genetic study of Common Hereditary Bowel Cancers in Hispania and the Americas", to Ian Tomlinson.

Authors contributions

VC: Conceptualization, data curation, formal analysis, methodology, writing original draft. PM: Conceptualization, formal analysis, writing original draft. PCH: Conceptualization, writing original draft. AC: Funding acquisition, project administration, supervision IQ: Data curation, investigation ARM: Funding acquisition, project administration, supervision, resources MS: Funding acquisition, project administration, supervision, visualization, writing – review & editing. All authors read and approved the final version.

References

Aaltonen LA, Johns L, Järvinen H, Mecklin JP and Houlston R (2007) Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res 13:356–361.

Alexander DH, Novembre J and Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664.

Boyle EA, Li YI and Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell 169:1177–1186.

Browning SR and Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.Am J Hum Genet 81:1084–1097.

Cantor RM, Lange K and Sinsheimer JS (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet 86:6–22.

Carvajal-Carmona LG, Cazier J-B, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J et al. (2011) Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and useof in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 20:2879–88.

Chakraborty R and Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85:9119–9123.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM and Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai H-J, Naqvi M, Phong A, Ung N et al. (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118:652–664.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Myers RM,Cavalli-sforza LL, Feldman MW and Pritchard JK (2009) The Role of Geography in Human Adaptation. PLoS Genet. 5: e1000500.

Darvasi A and Shifman S (2005) The beauty of admixture. Nat Genet 37:118–119.

Gabriel SB, Schaffner SFSF, Nguyen H, Moore JMJM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M et al. (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Galanter JM, Fernández-López JC, Gignoux CR, Barnholtz-Sloan J, Fernández-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras A V, Figueroa LU, Raska P et al. (2012) Developmentof a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet 8:e1002554.

González Burchard E, Borrell LN, Choudhry S, Naqvi M, Tsai H-J, Rodriguez-Santana JR, Chapela R, Rogers SD, Mei R, Rodríguez-Cintron W et al. (2005) Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research. Am J Public Health 95:2161–2168.

Haider S, Ballester B, Smedley D, Zhang J, Rice P and Kasprzyk A (2009) BioMart Central Portal–unified access to biological data. Nucleic Acids Res 37:W23–7.

Halder I, Shriver M, Thomas M, Fernandez JR and Frudakis T (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29:648–658.

Khoury MJ, Bedrosian SR, Gwinn M, Higgins JPT, Ioannidis JPA and Little J (2010) Human Genome Epidemiology. 2nd edition, Oxford University Press, New York.

Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D et al. (1991) Identification of FAP locus genes from chromosome 5q21. Science 253:661–665.

Lewontin RCC (1964) The interaction of selection and linkage II. Optimum models. Genetics 50:757–782.

Manta FSN, Pereira R, Caiafa A, Silva DA, Gusmão L and Carvalho EF (2013) Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. Ann Hum Biol 40:94–98.

Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM et al. (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 80:1171–1178.

Martinez-Fierro ML, Beuten J, Leach RJ, Parra EJ, Cruz-Lopez M, Rangel-Villalobos H, Riego-Ruiz LR, Ortiz-López R, Martínez-Rodríguez HG and Rojas-Martínez A (2009) Ancestry informative markers and admixture proportions in northeastern Mexico. J Hum Genet 54:504–509.

McKeigue PM (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76:1–7.

Moltke I and Albrechtsen A (2014) RelateAdmix: A software tool for estimating relatedness between admixed individuals. Bioinformatics 30:1027–1028.

Morton NE (2003) Genetic epidemiology, genetic maps and positional cloning. Philos Trans R Soc B Biol Sci 358:1701–1708.

Patel SR, Celedon JC, Weiss ST and Palmer LJ (2003) Lack of reproducibility of linkage results in serially measured blood pressure data. BMC Genet 4 Suppl 1:S37.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E and Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198–207.

Purcell S, Neale B, Todd Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Debakker P, Daly MJ et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

Qin H and Zhu X (2012) Power comparison of admixture mapping and direct association analysis in genome-wide association studies. Genet Epidemiol 36:235–243.

Rife DC (1953) Fingerprints as criteria of ethnic relationship. Am J Hum Genet 5:389–399.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516 – 1517.

Rojas W, Parra MV, Campo O, Caro MA, Lopera JG, Arias W, Duque C, Naranjo A, GarcíaJ, Vergara C et al. (2010) Genetic make up and structure of Colombian populations by meansof uniparental and biparental DNA markers. Am J Phys Anthropol 143:13–20.

Ros-Freixedes R, Battagin M, Johnsson M, Gorjanc G, Mileham AJ, Rounsley SD and Hickey JM (2018) Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. Genet Sel Evol 50:64.

Rubi-Castellanos R, Martínez-Cortés G, Muñoz-Valle JF, González-Martín A, Cerda-Flores RM, Anaya-Palafox M and Rangel-Villalobos H (2009) Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. Am J Phys Anthropol 139:284–294.

Salzano FM (2005) Evolutionary change - Patterns and processes. An Acad Bras Ciênc 77:627–650.

Salzano FM and Sans M (2014) Interethnic admixture and the evolution of Latin American populations. Genet Mol Biol 37:151–170.

Sans M (2000) Admixture studies in Latin America: from the 20th to the 21st century. Hum Biol 72:155–177.

Silva MCF, Zuccherato LW, Soares-Souza GB, Vieira ZM, Cabrera L, Herrera P, Balqui J, Romero C, Jahuira H, Gilman RH et al. (2010) Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: An application to populations of the state of Minas Gerais (Brazil). Genet Mol Res 9:2069–2085.

Skotte L, Jørsboe E, Korneliussen TS, Moltke I and Albrechtsen A (2019) Ancestry-specific association mapping in admixed populations. Genet Epidemiol 43:506–521. Stumpf MPH and Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13:1–8.

Teng B, Yang C, Liu J, Cai Z and Wan X (2016) Exploring the genetic patterns of complex diseases via the integrative genome-wide approach. IEEE/ACM Trans Comput Biol Bioinforma 13: 557-564.

The 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L,Gregersen PK et al. (2008) Analysis and Application of European Genetic Substructure Using 300 K SNP Information. PLoS Genet 4:e4.

Tishkoff SA and Verrelli BC (2003a) Patterns of human genetic diversity:implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340.

Tishkoff SA and Verrelli BC (2003b) Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. Curr Opin Genet Dev 13:569–575.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O et al. (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

Via M, Gignoux CR, Roth LA, Fejerman L, Galanter J, Choudhry S, Toro-Labrador G, Viera-Vera J, Oleksyk TK, Beckman K et al. (2011) History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. PLoS One 6:e16513.

Winkler CA, Nelson GW and Smith MW (2010) Admixture Mapping Comes of Age. Annu Rev Genomics Hum Genet 11:65–89.

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB et al. (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA 97:11354–11358.

Ziętkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R,Stoneking M, Tishkoff S, Batzer M et al. (1998) Genetic structure of the ancestral population of modern humans. J Mol Evol 47:146–155

Capítulo 3

Estructura poblacional

3.1. Estructura poblacional y estimaciones de parentesco en la población mexicana.

El siguiente capítulo consta de un artículo enviado y aceptado en marzo 2021 para ser publicado en *Annals of Human Genetics* como *Short Communication* (Colistro *et al.* 2021, ver Anexo II). El mismo no debía exceder las 2000 palabras, debe tener un máximo de dos tablas o figuras y hasta 15 referencias bibliográficas. Los autores del mismo son: Valentina Colistro, Augusto Rojas-Martínez, Ángel Carracedo, Ian Tomlinson, Luis Carvajal-Carmona, Raquel Cruz y Mónica Sans.

Considerando la compleja estructura poblacional evidenciada en el capítulo anterior, nos propusimos investigar el efecto del mestizaje y la consecuente estructuración poblacional sobre las estimaciones de parentesco en las muestras de México. En particular porque al estimar parentesco con el protocolo establecido por Anderson *et al.* 2010, el cual utiliza el programa

PLINK (Chang *et al.* 2015), detectamos un número desproporcionadamente alto de individuos emparentados. Esto nos hizo pensar que las estimaciones de parentesco, muy probablemente eran sesgadas, ya que dentro de los criterios de inclusión de las muestras colectadas para el proyecto, se incluyó no tener parientes de 1^{er} o 2^{do} grado que participaran del estudio.

Para ello analizamos los vínculos de parentesco entre la muestras de CHIBCHA de casos y controles, y aplicamos tres algoritmos distintos. Uno de los algoritmos es el cálculo estándar que suele aparecer en las guías de procedimientos de los análisis de GWAS y los otros son algoritmos específicamente diseñados para trabajar con poblaciones mestizadas. Los resultados muestran que hay una gran diferencia entre los algoritmos estándar de cálculo de parentesco y los diseñados a medida, siendo la ancestría genética el principal factor confusor. La principal conclusión de este análisis es la importancia de considerar la mezcla poblacional en etapas tempranas de los análisis genético poblacionales dado que podemos incurrir en sesgos en las estimaciones relacionadas al curado de los datos y los controles de calidad de los mismos, que nos conduzcan a posteriori, a resultados espurios.

En este trabajo también tuve un rol preponderante en todas la etapas que llevaron a la generación del borrados enviado, destacándose los análisis *in sillico*, el procesamiento bioinformático de las bases de datos, los análisis estadísticos, la visualización de los resultados y la redacción del manuscrito. Este trabajo fue realizado junto con la Dra. Mónica Sans, Dra. Raquel Cruz y Dr. Augusto Rojas-Martínez, quienes me guiaron en el los pasos a tomar que me condujeron a los resultados expuestos a continuación.

Population structure and relatedness estimates in a Mexican sample.

Running head: **Relatedness estimates in Admixed populations**

Colistro V.[1], Rojas-Martínez A.[2], Carracedo A.[3,4], CHIBCHA Consortium[5], Tomlinson I.[6], Carvajal-Carmona L.[7], Cruz R[*,4]., Sans M [*,8].

[1] Universidad de la República, Facultad de Medicina, Departamento de Métodos Cuantitativos, Montevideo,Uruguay.
[2] Escuela de Medicina y Ciencias de la Salud, Tecnológico de Monterrey, Monterrey, México
[3] Universidad de Santiago de Compostela, Centro Nacional de Genotipado (CEGEN), Spain
[4] Universidade de Santiago de Compostela, CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud
[5] Members of CHIBCHA Consortium are listed in the Acknowledgements section
[6] Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK
[7] Genome Center & Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, USA
[8] Universidad de la República, Facultad de Humanidades y Ciencias de la Educación, Departamento de Antropología Biológica, Montevideo, Uruguay

[*]Corresponding authors.

Keywords: Identity-by- descent, admixed populations and population stratification

Abstract

Population stratification (PS) is a confounding factor in genome-wide association studies and also, an interesting process itself. Latin American populations have mixed genetic ancestry, which may account for PS. We have analyzed the relatedness, by means of the identity-by-descent (IBD) estimations, in a sample of 1805 individuals and 1.006.703 autosomal mutations from a case-control study of colorectal cancer in Mexico. When using the recommended protocol for quality control assessment, 402 should have been removed due to relatedness. Our purpose was to analyze this value in the context of an admixed population. For that aim, we reanalyze

the sample using two software designed for admixed populations, obtaining estimates of 110 and 70 related individuals to remove. The results showed that the first estimation of relatedness was an effect of the higher Native American contribution in part of the data samples, being a confounding factor for IBD estimations. We conclude in the importance of considering PS and genetic ancestry in order to avoid spurious results, not only in GWAS but also in relatedness analysis.

Introduction

Population stratification has been widely studied as a confounding factor in genome-wide association studies (GWASs). For that purpose, it is important to accurately identify individuals with a high probability of being related to a certain degree, otherwise, the association will be biased. Usually, the determination of IBD (identity-by-descent) is used to detect relatedness. For samples that came from structured populations, it is not valid to assume population homogeneity. In spite of dealing with presumably unrelated samples, IBD/IBS (identity-by-descent and identity-by-state) estimates are essential in an early stage of any workflow for analyzing population structure.

Some authors have addressed the methodological issue of sameness due to shared ancestry as confounding in the determination of relatedness (Anderson and Weir, 2007; Wang, 2010). IBS is used in genetics to describe two identical alleles or two identical sequences of DNA, these alleles may be identical by chance or inherited from a recent common ancestor. IBD is defined as the proportion of 0, 1 or 2 identical by descent alleles between two individuals; the higher the estimate the greater the probability of relatedness. IBS represents the proportion of shared DNA segments, identical from the molecular point of view, but without sharing a common origin, or in which their common origin cannot be unequivocally determined (Forabosco *et al.*, 2005). Nevertheless, hidden or unrecorded relations may cause bias in the estimates.

Thornton *et al.* (2012), among others, have considered the problem of estimating relatedness in structured populations with admixed ancestry. However, having multiple source populations does not necessarily indicate population structure, in spite of multiple source populations can lead to more genetic diversity (Owings *et al.*, 2019).

The purpose of this paper is to analyze the relatedness in a case-control study for colorectal cancer in the Mexican population (unpublished data Colistro *et al.*, 2020), an example of admixed population, using different approaches. This concern arises from the first results for relatedness following the standard protocol (Anderson *et al.*, 2010).

Materials and Methods

We analysed 1805 samples from Mexico (929 cases and 876 controls), collected in three different locations: Mexico City, Monterrey and Torreón. We have chosen Mexican population as an example of admixed population, as its European contribution ranges from 40 to 62 % and the African, from 1 to 6 %, being the rest Native American (Salzano and Sans, 2014). All participants provided written informed consent for inclusion. The study was conducted in accordance with the Declaration of Helsinki. The protocol was approved by the ethics committees of each participating institution and the Federal Commission for Protection against Health Risks (COFEPRIS, Mexico), code CMN2012-001). Genotypes were obtained using two complementary arrays: Axiom Genome-Wide LAT 1 (Latino) Array and a Custom-designed Array, both from Affymetrix Axiom Genotyping Solutions. After SNP-level quality controls were applied, the resulting data set consisted of 1.006.703 autosomal SNPs uniformly distributed along the 22 autosomal chromosomes.

To assess the data quality of the analyzed samples we carried out the stepwise protocol described in Anderson *et al.* (2010). This protocol deals with the quality control (QC) of

genotype data from genome-wide association studies by using PLINK (Chang *et al.*, 2015) to carry out assessments of failure rate per-individual and per-SNP and to determine the degree of relatedness between individuals. At some point, the protocol proposes a principal component analysis (PCA) for the identification of individuals of divergent ancestry. In that matter, we first pruned typed SNPs in high linkage disequilibrium (LD) with r2 >0.1 and then we determined the PCs using EIGENSTRAT within smartpca (Price *et al.* 2006).

To replicate the results of the IBD obtained with PLINK, we used two software specifically designed to estimate relatedness in structured populations. KING (Manichaikul *et al.*, 2010) handles genotype data from GWASs or sequencing data and its algorithm considers the presence of population structure. Another program, REAP (Thornton *et al.*, 2012), estimates autosomal kinship coefficients and IBD sharing probabilities using SNP genotype data in samples with admixed ancestry. REAP also requires to include allele frequencies of SNPs in the ancestral populations and admixture proportions of each individual in the sample. So, for these estimations, we used data available in the 1000 Genomes Project (1000Genomes Consortium, 2010). Samples were selected to represent the parental populations of Mexico: African (N=176), European (N=174), and Asian (N=98), these last considered in substitution of Native Americans due to the scarcity of data about them, and the similarities because of the common origin. Estimation of individual admixture fraction was calculated with ADMIXTURE software version 1.3.1 (Alexander *et al.*, 2009), which considers a likelihood model. We assumed a three-population model with ancestry from African, European, and Native American populations. To choose the correct value of k we computed the cross-validation error over k, from 2 to 6. We found that $k = 3$ yielded the lowest cross-validation error (k_3=0.538).

Results

The estimation of IBD using PLINK, showed an oddly high number of related samples, 402 samples to remove (23 % of samples, IBD>0.1875) (Table 1). This estimation is extremely unlikely since being a relative of another participant was an exclusion criterion during the sample collection process in Mexico. Most of the related samples formed blocks of individuals related all to all, rather than one to one (pairs) or one to a few. This made us think that some variable was generating false closeness. The samples reported as related were different when analysing each array separately (92 in the CUSTOM array and 373 in the Latino array). To understand which factors were being confounded with relatedness we evaluated several variables: sampling location, case/control status, ancestry estimation, sex, genotype batch and significant principal components. The admixture analysis stratified by case/controls showed no significant differences between groups. However, cases tend to present greater Caucasian ancestry(41.27 % vs 38.5 % in controls), while controls presented an increased Asian ancestry (55.9 % versus 53.23 %). The African ancestry was the smallest in both groups, 5.6 % in cases and 5.02 % in controls. The mean global ancestry in the whole dataset was 54.3 % Asian, 40.4 % of Caucasian and 5.3 % of African. We detected that Asian ancestry was highly correlated PC1, and pairs of samples with high IBD positioned at the very end of the distribution (higher Asian ancestry and lower eigenvalue for PC1) (Fig1).

Lately, we assumed structured samples and used KING and REAP software. Both consider the structure of the samples due to shared ancestry to avoid identifying samples of the same ancestry as related. When running KING, 110 samples have to be removed (Table 1). The results of REAP showed consistency with KING, but in this case, only 73 samples to remove were identified (Table 1). To assure comparability, we used the IBD thresholds proposed in KING to determine the degree of relatedness on REAP results.

Tabla 3.1: Pairs of samples identified as related by three different software, according to the degree of relatedness.

Relationship	PLNIK	REAP	KING
Identical twins	10	10	10
1st degree	55	59	52
2nd degree	24	32	29
3rd degree	313	17	19
Total	402	73	110

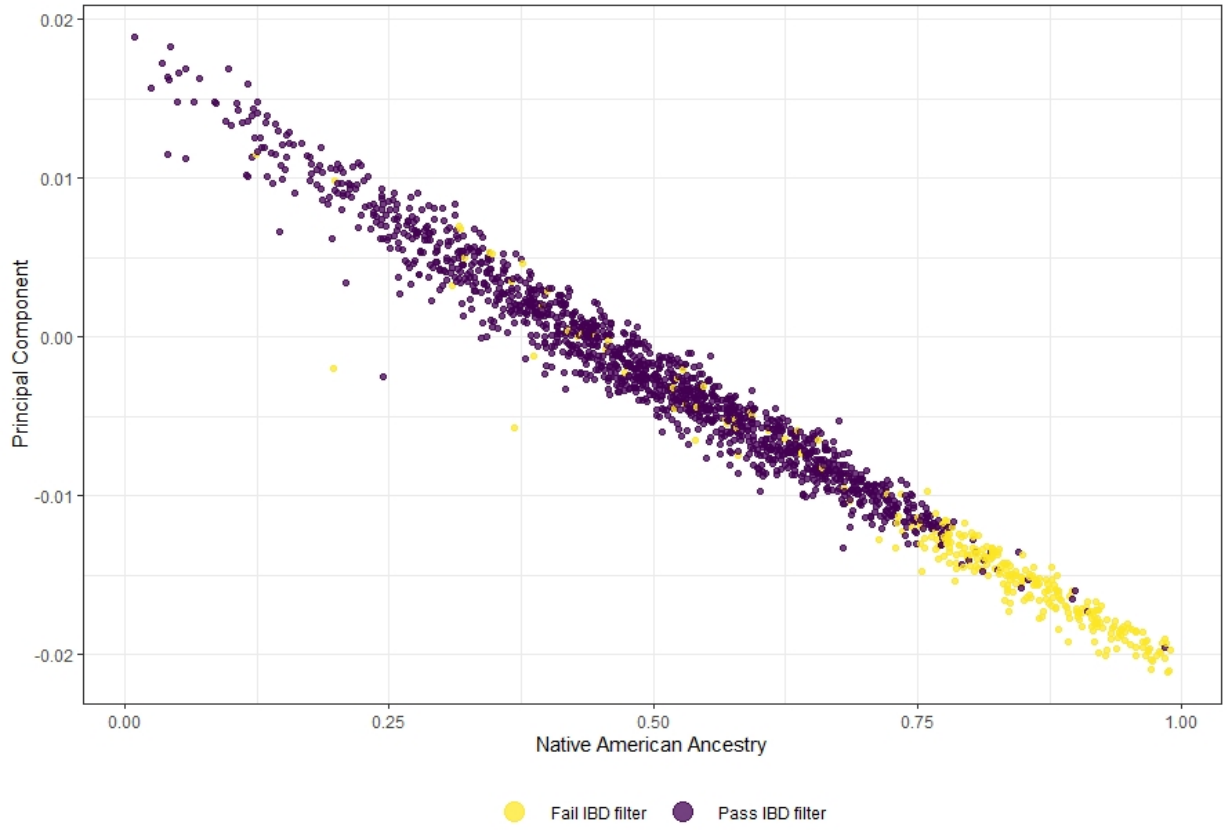

Figura 3.1: Principal Component 1 against Native American ancestry. Pairs of samples with high IBD (>0.1875) are spotted in yellow.

Discussion and conclusions

Genetic diversity is lower in Native American populations related to other continental regions, measured by heterozygosity (Wang *et al.*, 2007). On the other hand, Mexicans have a high degree of differentiation between populations explained by high degrees of isolation, measured by FST. Moreover, those populations appear as discrete units, with scarce gene flow (Moreno-Estrada *et al.* 2014). We observed that higher IBD values are related to increased Native American ancestry when using PLINK.We postulate that the isolation that leads to a high differentiation among Mexican Natives is a confounding factor for IBD estimation due to deep sub-structuration, and each ethnic group can be taken as a familiar group as happens with software not developed to deal with admixed samples. With Mexican mestizo samples, we have shown that the software confounded Native-American ancestry with relatedness, overestimating the relatedness among samples. This has practical implications when studying admixed populations as we have shown that exclusion of samples based on measures of relatedness may produce biased results. Lastly, differences found between both arrays can be explained because the Latino Array was designed to highlight differences among parental populations.

Acknowledgements: CHIBCHA Consortium (study of hereditary cancer in Europe and Latin America), Ian Tomlinson (University of Edinburgh, UK); Luis Carvalal-Carmona (University of California, Davis, USA), Ma. Magdalena Echeverry de Polanco, Mabel Elena Bohórquez, Rodrigo Prieto, Angel Criollo, Carolina Ramírez, Ana Patricia Estrada, Jhon Jairo Suárez (Universidad del Tolima, Colombia); Augusto Rojas Martinez, Rocío Ortiz Lopez (Tecnológico de Monterrey, Mexico); Silvia Rogatto, Samuel Aguiar Jnr, Ericka Maria Monteiro Santos (São Paulo State University, Botucatu, Brazil); Monica Sans, Valentina Colistro, Pedro C. Hidalgo, Patricia Mut (University of the Republic, Uruguay); Angel Carracedo, Clara Ruiz Ponte, Ines Quntela Garcia (-University of Santiago de Compostela, Spain); Sergi Castellvi-Bel

(University of Barcelona, Barcelona, Catalonia, Spain); Manuel Teixeira (Portuguese Oncology Institute, Portugal).

Conflict of interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Data availability statement: The dataset generated and analyzed during the current study are available from corresponding authors on reasonable request.

Authors contribution: Study design: LCC, IT; Conceptualization: VC, AR, IT, LC, MS; Data curation: RC; Formal Analysis: VC, AC, RC; Methodology: VC, RC, MS; Funding acquisition: ARM, AC, CC, IT, LCC; Visualization: VC; Project Administration: ARM, AC, CC, IT, LCC; Supervision: MS; Writing original draft: VC, ARM, MS.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061–1073. https://doi.org/10.1038/nature09534

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics, 176(1), 421–440. https://doi.org/10.1534/genetics.106.063149

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. Nature Protocols, 5(9), 1564–1573.https://doi.org/10.1038/nprot.2010.116

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

Forabosco, P., Falchi, M., & Devoto, M. (2005). Statistical tools for linkage analysis and genetic association studies. Expert review of molecular diagnostics, 5(5), 781–796.https://doi.org/10.1586/14737159.5.5.781

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford, England), 26(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., Via, M., ... Bustamante, C. D. (2014). Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science (New York, N.Y.), 344(6189), 1280–1285. https://doi.org/10.1126/science.1251688l

Owings, A. C., Fernandes, S. B., Olatoye, M. O., Fogleman, A. J., Zahnd, W. E., Jenkins, W. D., Malhi, R. S., & Lipka, A. E. (2019). Population Structure Analyses Provide Insight into the Source Populations Underlying Rural Isolated Communities in Illinois. Human biology, 91(1), 31–47.https://doi.org/10.13110/humanbiology.91.1.05

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics, 38(8), 904–909. https://doi.org/10.1038/ng1847

Salzano, F. M., & Sans, M. (2014). Interethnic admixture and the evolution of Latin American populations. Genetics and molecular biology, 37(1 Suppl), 151–170.https://doi.org/10.1590/s1415-47572014000200003

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. American journal of human genetics, 91(1), 122–138. https://doi.org/10.1016/j.ajhg.2012.05.024

Wang J. (2011). Unbiased relatedness estimation in structured populations. Genetics, 187(3), 887–901. https://doi.org/10.1534/genetics.110.124438

Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A. M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J. A., Freimer, N. B., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., Tsuneto, L. T., ... Ruiz-Linares, A. (2008). Geographic patterns of genome admixture in Latin American Mestizos. PLoS genetics, 4(3), e1000037. https://doi.org/10.1371/journal.pgen.1000037

Capítulo 4

Estudio de asociación genómico con el CRC

4.1. Estudio de asociacón genómico a cáncer colorrectal en individuos mexicanos sugiere nuevas variantes de riesgo.

Este capítulo se presenta en formato de artículo a ser enviado para su consideración por parte de la revista *Frontiers in Oncology*. Los autores del artículo son: Valentina Colistro, Raquel Cruz, Inés Quintela, Clara Ruiz, Angel Carracedo, Pedro Luna-Pérez, Irma S. García-Gonzalez, Carlos Murillo-Martinez, Rocío Ortiz-López, Yolanda Jaramillo-Rodríguez, Pablo Ruiz-Flores, Edmundo Castelan-Maldonado, Juan F. González-Guerrero, Sergio Cárdenas-Cadena, Nidia K. Moncada-Saucedo, Jorge Haro-SantaCruz, Fátima M.Alvarado-Monroy, Mónica Sans y Augusto Rojas-Martínez; todos investigadores del proyecto CHIBCHA.

El trabajo consta de un estudio de asociación de regiones genómicas asociadas a CRC. Para el mismo se contó con con 831 casos y 881 controles genotipados para 1,006,703 polimorfismos autosómicos. Los controles de calidad de las muestras genotipadas están detallados en Anderson *et al.* 2010, con la consideración del cálculo de IBD detallado en el capítulo precedente. El GWAS se realizó aplicando un modelo de regresión en el cual se incluyeron variables confusoras para evitar asociaciones espúreas. Estos análisis se realizaron usando principalmente el *software* PLINK, ampliamente utilizado en análisis genéticos. Entre las variables confusoras se consideró la estructuración poblacional y para ellos se realizó un análisis de componentes principales utilizando el *software* Eigenstrat (Patterson *et al.*, 2006; Price *et al.*, 2006), para estos cálculos se consideraron muestras extraídas de 1000G con ancestrías predominantemente europea, africana y muestras con ancestría nativo-americana publicadas previamente (Mao *et al.* 2007).

El GWAS arrojó un SNP fuertemente asociado (*rs35797542*; valor p $< 5\times10^{-8}$), localizado en la región 8q24.22, y otros 16 SNPs mostraron una asociación sugerente (5×10^{-8} < valor-p $< 1 \times 10^{-5}$). Estos 17 SNPs fueron referidos a los genes más cercanos, lo cual representó 12 genes, y dentro de los límites génicos se realizó la imputación de genotipos con el fin de aumentar la densidad de variantes. Para el proceso de imputación se consideraron genomas de 2,504 muestras pertenecientes a 26 poblaciones mundiales (datos extraídos de la plataforma 1000G). Con los SNPs genotipados y los imputados, se realizó un SKAT, el cual estima la asociación de los genes con el CRC, considerando el efecto combinado de todos los SNPs dentro de los límites génicos. Éste análisis detectó asociación en 5 genes: *ZFAT, LRRC36, PLEKHG4, ATP6V0D1* y *KCTD19*. El gen *ZFAT* es donde está ubicado *rs35797542* y no ha sido previamente reportado como asociado a CRC; este gen codifica una proteína de unión al ADN mediante dedos de zinc y funciona como un regulador transcripcional involucrado en la apoptosis y la supervivencia

celular y ha sido descrita como vinculada a la enfermedad autoinmune tiroidea. Los restantes cuatro genes asociados (*LRRC36, ATP6V0D1, KCTD19*, y *PLEKHG4*) están todos ubicados en la misma citobanda: 16q22.1. El gen *LRRC36* codifica para un proteína rica en repetidos de leucina y ha sido previamente asociada a obesidad y distribución corporal de grasa. Ninguno de los genes *KCTD19 y PLEKHG4* fue descrito en un GWAS previamente. *KCTD19* codifica el tetrámero de canales potasio, no está claro su vinculación con una enfermedad, pero su localización en el genoma se superpone con la región promotora del gen *LRRC36*, lo cual podría explicar detectar ambos genes como asociados a CRC. Lo que respecta a *PLEKHG4*, también codifica una porteína a cual cumple funciones de intercambio de nucleótidos de guanina y cree que puede desempeñar un papel en la señalización intracelular y la dinámica del citoesqueleto en el aparato de Golgi. Este estudio evidencia la importancia de realizar estudios en poblaciones latinoamericanas y su potencialidad en identificar nuevas regiones candidatas a colaborar en la etiología y desarrollo del CRC. Tiene la limitante del tamaño muestral y de la escasez de genomas de parentales nativas específicas de la población mexicana. Sería bueno poder contar con muestras de casos y controls de CRC de otras poblaciones mestizadas latinoamericanas para poder replicar los resultados.

En este trabajo también tuve un rol principal en las diversas etapas del trabajo. Esto incluye tanto el curado de los datos, como el diseño del trabajo, los análisis *in silico* y la redacción del manuscrito. Este trabajo fue un esfuerzo conjunto de varios grupos de investigación involucrados en el proyecto CHIBCHA y en particular el Dr. Augusto Martínez y la Dra. Mónica Sans, fueron quienes orientaron el trabajo.

A genome-wide association study of colorectal cancer in Mexican mestizos suggest novel common tumor-risk variants

Colistro V.[1], Cruz R.[2], Quintela I.[3], Ruiz C.[3], Carracedo A. [2,3], Luna-Perez P.[4], Garcia-Gonzalez IS.[5], Murillo-Martinez C.[6], Ortiz-Lopez R.[7], Jaramillo-Rodriguez Y.[8], Ruiz-Flores P.[9], Castelan-Maldonado E.[10], Gonzalez-Guerrero JF.[11], Cardenas-Cadena S.[12], Moncada-Saucedo NK.[12], Haro-Santa Cruz J.[13], Alvarado-Monroy FM. [7], Tomlinson I. [14], Carvajal-Carmona L.[15], CHIBCHA Consortium [16], Sans M [17,*], Rojas-Martínez A. [18,*],

[1] Universidad de la República, Facultad de Medicina, Departamento de Métodos Cuantitativos, Montevideo,Uruguay.
[2] Universidade de Santiago de Compostela, CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud
[3]Grupo de Medicina Xenómica, Centro Nacional de Genotipado (CEGEN-PRB3-ISCIII). Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[3]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Oncology Department. Mexico City, Mexico.
[5]Instituto Mexicano del Seguro Social, Medical Unit of High Specialties 25, Surgical Oncology Service Monterrey, Mexico.
[6]Instituto Mexicano del Seguro Social, Centro Médico Nacional Siglo XXI, Central Blood Bank. Mexico City, Mexico.
[7] Tecnológico de Monterrey, Escuela de Medicina y Ciencias de la Salud. Monterrey, México.
[8]Instituto Mexicano del Seguro Social, Medical Unit of High Specialties 71 and Unit of Family Medicine 16. Torreon, Mexico.
[9]Universidad Autónoma de Coahuila at Torreon, University Hospital and School of Medicine. Torreon, Mexico.
[10]Instituto Mexicano del Seguro Social, Medical Unit of High Specialties 25, Pathology Service. Monterrey, Mexico.
[11]Universidadde Nuevo León, University Hospital, Center against Cancer. Monterrey, México.
[12]Universidad de Nuevo León, University Hospital, School of Medicine. Monterrey, México.
[13]Universidad Autónoma de Coahuila at Torreon, School of Medicine. Torreon, Mexico.
[14]Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK
[15]Genome Center & Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, USA
[16] Members of CHIBCHA Consortium are listed in the Acknowledgements section
[17] Universidad de la República, Facultad de Humanidades y Ciencias de la Educación, Departamento de Antropología Biológica, Montevideo, Uruguay
[18] Escuela de Medicina y Ciencias de la Salud, Tecnológico de Monterrey, Monterrey, México

*Corresponding authors.

Abstract: Genome-wide association studies (GWAS) for colorectal cancer (CRC) have detected high-risk genetic variants associated with CRC in several ethnic groups, but Latin American communities are still underrepresented. The aim was to identify variants related to CRC in an admixed Latin American population. **Methods:** The study was performed in 831 cases and 881 controls from Mexico, who were genotyped for 1,006,703 autosomal SNPs. Logistic regression was carried out including covariants, such as sex, age and genetic ancestry. Lastly, we performed a sequence-kernel association test (SKAT) to consider the joint effect of several SNPs lying in genes. **Results:** One chromosomal region reached genome-wide significance level ($p < 5 \times 10^{-8}$): rs35797542 - 8q24.22 and 16 variants reached borderline statistical significance ($p < 1 \times 10^{-5}$). SKAT analysis detected 5 candidate genes associated with CRC, where only one (*ZFAT* located in 8q24.22) was also detected in the GWAS. This gene was not previously reported in association with CRC. Other 4 genes (*LRRC36, ATP6V0D1, KCTD19,* and *PLEKHG4*) were highly suggestive for further association and functional studies. **Conclusions:** We found 1 SNP (rs35797542) and 1 gene (*ZFAT*) associated with CRC, and 4 other genes These signals may contribute to enrich the panoply of genes involved with CRC. Further analyses remain to be done to validate the associations in other Latin American populations. This study highlights the importance of conducting GWAS in poorly explored admixed populations.

Keywords: Single Nucleotide Polymorphisms; Sequence-kernel association test; GWAS; Mexico; Colorectal Cancer

Running title: GWAS of colorectal cancer in Mexicans

1. Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide and it is the fourth leading cause of cancer deaths worldwide considering both sexes. In Mexico, CRC is the third most common cancer and the most common cause of cancer-related death considering both sexes. CRC incidence rate has steadily increased in the past 20 years in Mexico, going from 8.28 new cases per 100,000 in 1996 to 21.9 new cases per 100,000 in 2017 (1,2).

The etiology of CRC is not well established and has few avoidable risk factors. Several studies have shown that common inherited single nucleotide polymorphisms (SNPs) can increase CRC risk. Previous genome-wide association studies (GWAS) identified common variants associated with CRC. To date, 58 susceptibility alleles across 37 regions associated with $p < 5 \times 10^{-8}$ have been identified (3), most of these alleles were identified in Caucasian populations. However, those SNPs accounts for $<5\,\%$ of the genetic risk of CRC and it will be very important when the effects of those already detected variants are added to those as-yet undetected CRC risk SNPs. Evidence to date suggests that many CRC genes are associated with low relative risks (4).

GWAS have determined that common variants only explain a small proportion of the phenotypic variance for most complex traits studied to date. Manolio et al. (5) suggested that rare variants could contribute substantially to the missing heritability. Several GWAS studies conducted in admixed populations evidenced the importance of considering populations with a history of recent admixture in order to detect polymorphisms undetectables in non admixed populations. These populations have longer linkage disequilibrium (LD) blocks than non-admixed populations such as Europeans or Africans (6–8). Genetic epidemiological studies in Latin American (LA) populations are thus very promising, not only because of their historical importance,

but also because they can provide new important insights into disease. Ancient gene pools, such as those from Africa, have a greater amount of overall variation and a finer LD structure between markers (9). Maximum ability to differentiate populations comes from genetic markers with large frequency differences among parental populations for admixed samples. In combination, the unique LD structure and allele frequency spectrum of admixed populations may assist in localizing association signals (10,11).

GWAS studies enable identifying loci independently associated to CRC, but in order to establish the association of a certain gene, the combined effect of several SNPs within the limits of the gene should be considered. Some variants, common or rare, may have unpredictable frequencies in a particularly studied population. Groupwise association tests may contribute to upweight the contribution of rare variants and down-weight the contribution of common variants.

The present study aimed at identifying variants conferring genetic risk to CRC in the population of Mexico, a Latin American country with a large admixed population, by performing a GWAS to investigate the contribution of SNPs to CRC. We also tested the association of the nearest gene to the SNPs detected, in combination with the joint effect of several SNPs (genotyped and imputed) lying within the limits of the gene.

2. Materials and Methods

Study subjects

Cases were recruited in three different locations within Mexico, Mexico City (74.6 %, from Centro Médico Nacional Siglo XXI of the Mexican Social Security Institute -IMSS-), Monterrey (22.7 %, from UMAE 25, IMSS and the University Hospital of the Universidad Au-

tónoma de Nuevo León) and Torreón (2.7 %, from the UFM 16 IMSS, the UMAE 71 IMSS, and the University Hospital of Torreón), from 2010 to 2012. Control samples were collected from blood banks located at the same medical centers in Mexico City, Monterrey, and Torreón (72.6, 25.9 and 1.5 %, respectively). A total of 1,943 samples were collected (1,123 males, 820 females), being 960 cases and 983 controls. DNA of all samples were extracted from the peripheral blood using Wizard® Genomic DNA Purification columns (Promega, Inc. Madison, WI). All participants provided written informed consent for inclusion, before their participation. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the ethics committees of each participating institution (Ethics Committee of the University Hospital, Universidad Autónoma de Nuevo León code BI10-003, the National Commission of Scientific Research of the Mexican Social Security Institute code R-2012-785-032), and the Federal Commission for Protection against Health Risks (COFEPRIS), code CMN2012-001, Mexico.

Genotyping and Quality Control

Samples were genotyped using two complementary arrays: Axiom Genome-Wide LAT 1 (Latino) Array and a Custom-designed Array, both from Affymetrix Axiom Genotyping Solutions. The former was designed to maximize coverage of common and rare disease-associated alleles in Latin American populations that have genetic contributions from European, Native American and African ancestries. The latter was specifically designed for this study. being the SNPs selection based on regions previously detected as associated with CRC. SNP calling in both arrays was done following Affymetrix best practice workflow, which includes the Genotyping Console Software in combination with SNPolisher.

Anderson et al. (12) protocol was followed to assess data quality. A total of 1,169,944 SNPs (387,948 from the Custom Array and 781,996 from the Latino Array) was obtained. SNPs in Chromosome Y and mitochondrial DNA were excluded. Individuals were filtered out based on discordant sex information, missing rate (<90 %), outlier heterozygosity, related individuals and divergent ancestry. Markers were filtered based on differences in call rate between cases/controls and missing rate per marker (<2 %). Allele frequencies of variants were verified and only variants with MAF greater than 1 % remained (Figure S2). Quality control stages were carried out using PLINK v1.9 (13,14)

In order to accurately identify duplicated or related individuals we used REAP software version 1.2 (15) This program estimates relatedness in structured populations with admixed ancestry. Pairs of individuals were considered closely related if the estimated kinship coefficient between them was 0.1. Failing individuals in every single step were removed for the downstream analysis. Once the quality controls were finished the resulting data set consisted of 1,712 samples (831 cases and 881 controls) and 1,006,703 autosomal SNPs. Variants were uniformly distributed along the 22 autosomal chromosomes as shown in Figure S1. Allele frequencies of variants were verified and only variants with minor allele frequency (MAF) greater than 1 % remained (Figure S2).

Population stratification

We implemented a principal component analysis (PCA) to assess genetic ancestry. First we pruned genotyped SNPs in LD (r2 >0.1) using PLINK. With the remaining data, we determined the PCs using EIGENSTRAT v6.0.1 within SmartPCA (16,17). The PCA was run twice, a first run on Mexican case/control samples in combination with parental populations, European and African samples from 1000 Genomes (CEU and YRI, respectively) and the Native

American samples from Mao et al. (18). These Native American samples are mainly composed by Nahuatl and Maya subjects from Mexico, and a South American sample comprising Aymaras/Quechuas (Bolivia) and Quechua (Peru). (Figure 1A) Subsequently, another run was performed only with study samples, to generate PCs for global ancestry adjustment in association analyses. We used the Tracy-Wisdom statistics to evaluate the statistical significance of each PC identified by PCA. The distribution of the PCs was similar in cases and controls. (Figure 1B).

Association

We performed GWAS tests using multivariate logistic regression, as implemented in PLINK, using an additive genetic model adjusted for: age, sex and PCs. To calculate LD, we calculated r^2 in the controls in our dataset using PLINK. We then performed conditional analyses by entering the most significant SNPs in the model as a covariate, in addition to PCs. Correlation among PCs was estimated and two PCs showed to be highly correlated between them and with PC1 so they were excluded. The two most informative non-correlated PCs were used to adjust for population stratification in the association tests. Genome-wide inflation was evaluated by estimating λ.

Imputation

In order to increase SNP density along associated genome regions, imputation of unobserved genetic polymorphisms was performed. The process was carried out based on SNPs passing quality controls, which were phased using Eagle v2.4 (19). Imputation was conducted using Minimac3 (20) and phase 3 of the 1000 Genomes Project (comprising 2,504 individuals from 26 populations worldwide) as the reference panel. Imputed SNPs with deviations from Hardy-Weinberg equilibrium were filtered out of downstream analyses.

Sequence kernel association test

SKAT (21) tests the combined effects of multiple variants (common and rare) in a region on a phenotype. In this study, regions were defined considering those associated SNPs lying within the limits of a coding sequence. As mentioned above, we considered SNPs strongly associated ($p < 5 \times 10^{-8}$) but also SNPs with highly suggestive CRC associations ($p < 1 \times 10^{-5}$). From those regions, both genotyped and imputed SNPs were considered together to test association to CRC.

3. Results

A total of 231 samples were removed. Fourty-eight samples because of discordant sex information, 45 individuals because elevated proportion of missing genotypes (>0.01), 41 due to heterozygosity rate beyond ± 3 standard deviations from the mean. Besides, seventy-two samples were eliminated from further analysis because of their relationship estimated by kinship coefficients for pairs of individuals (22). Lastly, 25 samples were eliminated due to African Ancestry >0.20. Alleles with MAF <1 %, which represented 9.8 % of the variants (619 showed no polymorphism, MAF = 0) and autosomal SNPs showing significant deviation from the Hardy-Weinberg equilibrium in controls (p-value <10-3) were eliminated. No marker was removed due to different genotype call rates between cases and control. After quality control, 1,712 adults (1,012 males, 700 females, constituting 831 cases and 881 controls) and 1,006,703 SNPs were available for downstream analyses.

Population Stratification

As shown in Figure 1, Mexican samples mainly lie between the Caucasians and Native Americans, coherently with the demographic history of the Mexican population. Both, Mexican

cases and controls, are plotted far from the African samples. Moreover, when observing PC3, the Native American samples split into two different subsamples (Fig. 1 b and c). As stated previously, Native American samples comprise a Mesoamerican sample of Maya and Nahua from Mexico, and a South American sample comprising Aymara/Quechua (Bolivia) and Quechua (Perú) subjects.

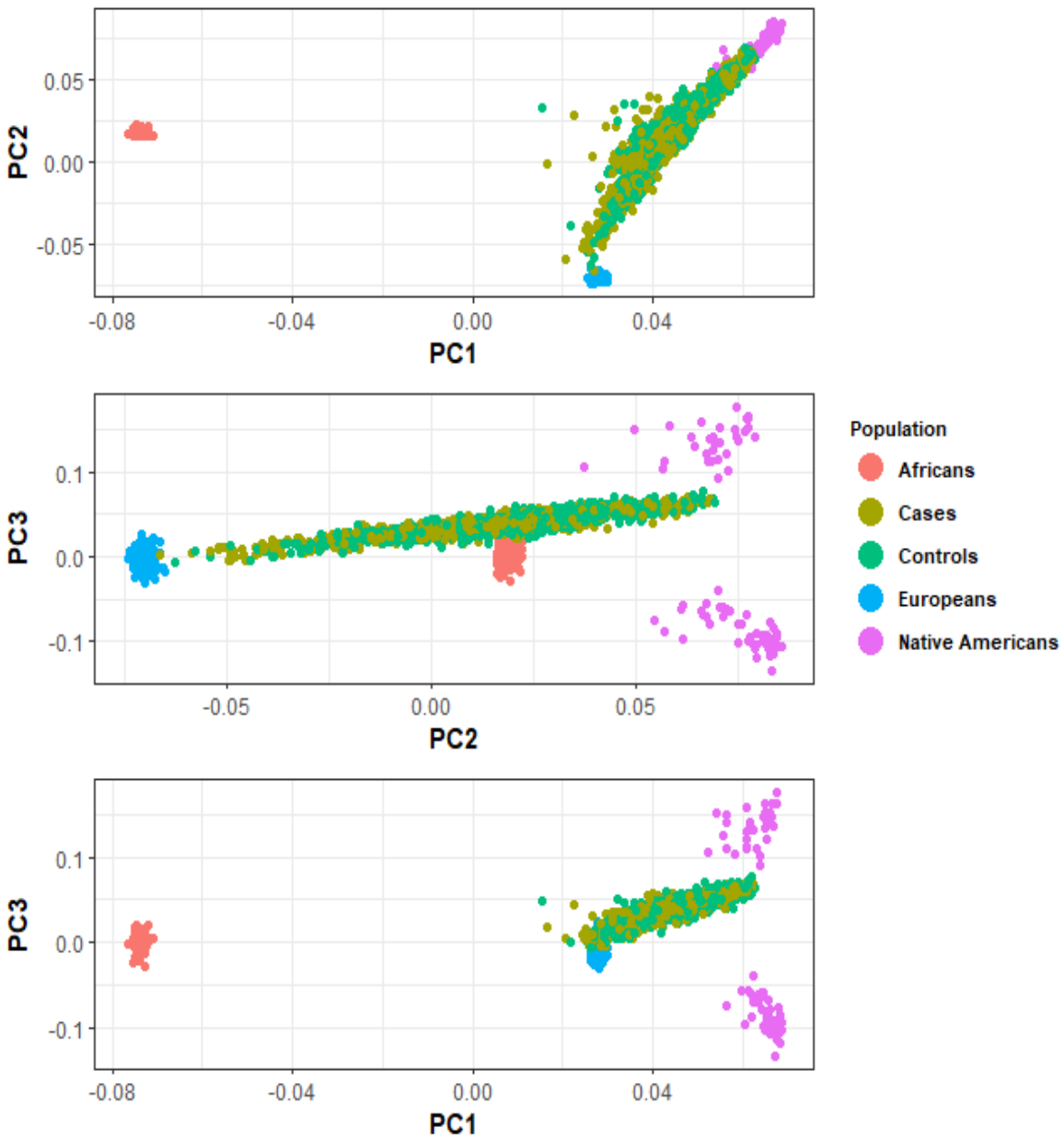

Figure 1. Principal components analysis.S These figures show the clustering results using principal components analysis implemented by the Eigensoftware with genome-wide independent autosomal SNPs in CRC case/control samples and individuals from European and African ancestry from 1000 (Phase 3) and Native Americans. a) PC1 versus PC2, b) PC2 versus PC3 and c) PC1 versus PC3

Association test

Following a PCA adjustment for ancestry, sex and age, the genomic inflation factor (λ) was 1.03, indicating a low probability of false-positive associations as a result of population structure. The most strongly associated SNP (rs35797542) was part of the SNP set of the Custom Array and was found on genomic region 8q24.22; the remaining 16 variants that reached borderline statistical significance were found on: chromosome 2 (rs7589473 on q24.4 and rs1399958 on q31.1), chromosome 6 (rs16868695 on p21.31), chromosome 7 (rs255167 and rs73087778 on p14.3), chromosome 12 (rs11173904 on q14.1), chromosome 13 (rs1750424 on q33.1) and chromosome 16 (rs9922476, rs16957289, rs16957304, rs7205526, rs8047080, rs8052655, rs9922085, rs11860295, rs3868142 on q22.1)(Figure 2). The list of overlapping genes and nearest upstream/downstream genes was obtained from the Ensembl database (23) (Table 2).

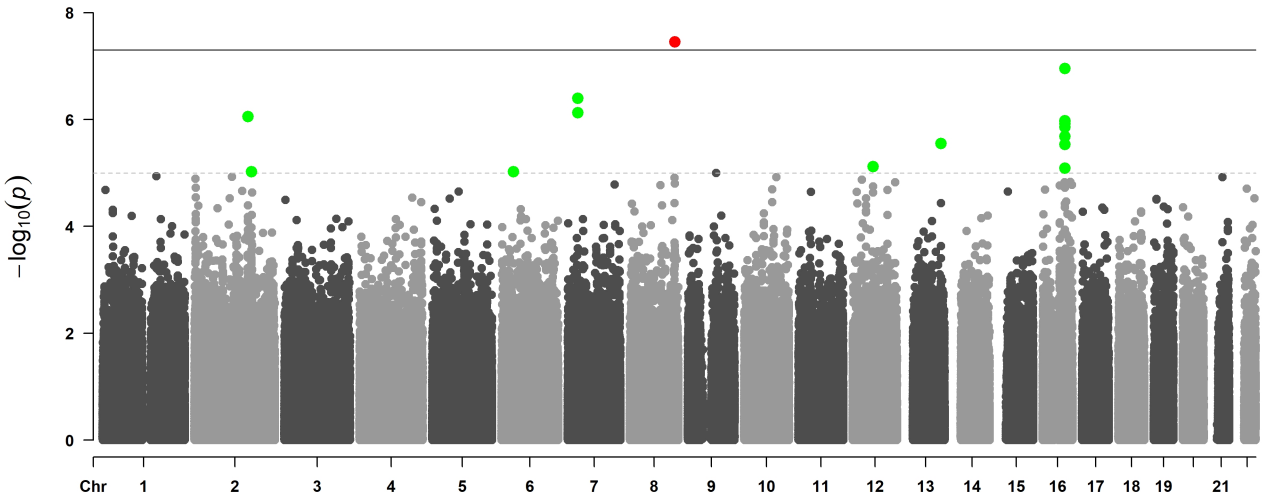

Figure 2. Manhattan plot of association p values. p values of 1.006.703 autosomal SNPs are shown. The solid horizontal line represents the genomewide significance threshold of $p = 5.0 \times 10^{-8}$ and the dashed line $p = 1.0 \times 10^{-6}$.

Imputation

We imputed genotypes within the limits of genes where the 17 SNPs were located. Among those 17 variants (16 borderline significant and 1 highly significant), 12 lie within genes while the remaining 5 lie in intergenic regions. When considering the location of the intragenic SNPs, 7 correspond to intronic variants, 1 to 5'UTR variants and 4 to exonic variants (all four with non synonymous consequence to transcript). It is worth to notice that 8 out of those 12 variants were located on three genes of cytoband 16q22.1: four SNPs belong to the *LRRC36*, two belong to the *PLEKHG4* and two belong to *KCTD19*. The remaining 4 variants were all located in different genes (Table 1).

Table 1. SNPs associated with CRC risk. SNPs ids, position, risk allele, OR, nearest gene and consequence to transcript of the significant SNPs ($p < 1.0 \times 10^{-5}$).

SNP	Cytoband	Position (bp)	Risk Allele	OR (IC 95 %)	p-value	Nearest Gene	Consequence to transcript
rs7589473	2q24.2	161369139	A	1.62 (1.33-1.96)	8.784×10^{-7}	*RBMS1*	upstream
rs1399958	2q31.1	172525144	T	1.54 (1.27-1.86)	9.505×10^{-6}	—	downstream
rs16868695	6p21.31	35214884	T	2.63 (1.71 - 4.03)	9.462×10^{-6}	*SCUBE3**	intron
rs255167	7p14.3	30770748	A	3.30 (2.01 - 4.34)	4.025×10^{-7}	*INMT**	intron
rs73087778	7p14.3	30823918	G	6.65 (3.14 - 12.7)	7.463×10^{-7}	*FAM188B**	intron
rs35797542	8q24.22	135464721	C	1.92 (1.52 - 2.42)	3.515×10^{-8}	*ZFAT*	downstream
rs11173904	12q14.1	61742195	T	1.86 (1.42 - 2.44)	7.643×10^{-6}	*SLC16A7*	downstream
rs1750424	13q33.1	103865749	T	1.59 (1.31- 1.93)	2.806×10^{-6}	*SLC10A2*	upstream
rs11860295	16q22.1	67316234	T	2.88 (1.86 - 4.46)	2.062×10^{-6}	PLEKHG4*	missense
rs3868142	16q22.1	67320223	A	2.88 (1.86 - 4.46)	2.083×10^{-6}	PLEKHG4*	missense
rs16957289	16q22.1	67325711	T	4.07 (2.3 - 7.20)	1.397×10^{-6}	*KCTD19**	missense
rs16957304	16q22.1	67334969	G	3.43 (1.99 - 5.91)	8.181×10^{-6}	*KCTD19**	intron
rs9922085	16q22.1	67397580	C	3.31 (2.04 - 5.37)	1.104×10^{-6}	*LRRC36**	UTR-5
rs8047080	16q22.1	67402588	G	2.97 (1.91 - 4.62)	1.203×10^{-6}	*LRRC36**	intron
rs8052655	16q22.1	67409180	A	3.40 (2.08 - 5.56)	1.055×10^{-6}	*LRRC36**	missense
rs7205526	16q22.1	67410583	T	3.43 (2.17 - 5.41)	1.108×10^{-7}	*LRRC36**	intron
rs9922476	16q22.1	67478924	G	3.17 (1.95 - 5.14)	2.911×10^{-6}	*ATP6V0D1**	intron

*SNPs are located within the limits of the genes

Genotype data for a total of 15,444 SNPs were available after imputation within the genes limits, including both, fully imputed variants, as well as genotyped SNPs with missingness

<10 %. As part of the imputation quality control, a subset of genotyped SNPs were masked and underwent the complete imputation process. The genotyped frequency was compared with the imputed frequency of the subset. The analysis demonstrated high-quality imputation with mean differences not greater than 0.4 %. We filtered out non bi-allelic and non-SNP imputed variants.

Sequence-kernel association test

SNPs from Table 1 were referred to the gene where they lie on or to the nearest gene. This represented 12 different genes, which were analysed using SKAT considering imputed and genotyped SNPs within the limits of those loci. Table 2 summarizes the variants, their location, the associated genes and the results of the SKAT. Five genes showed association with CRC when considering the combined effects of several genotyped and imputed SNPs in that same gene (*ZFAT*, *ATP6V0D1*, *LRRC36*, *KCTD19* and *PLEKHG4*). The *ZFAT* locus is highly suggestive because it is the gene where rs35797542 lies on, thus we confirmed not only SNP association, but also gene association (Table 2).

Table 2. SKAT results of the analyses, genes, number of markers tested and p-values.

Gene	CHR	Cytoband	SKAT p-value	Markers Tested
ATP6V0D1	16	q22.1	1.65×10^{-7}	228
LRRC36	16	q22.1	3.51×10^{-7}	289
KCTD19	16	q22.1	5.52×10^{-6}	173
ZFAT	8	q24.22	6.07×10^{-5}	813
PLEKHG4	16	q22.1	1.37×10^{-3}	47
SCUBE3	6	p21.31	2.91×10^{-1}	289
RBMS1	2	q24.2	4.98×10^{-1}	1630
FAM188B	7	p14.3	5.07×10^{-1}	494
SLC10A2	13	q33.1	6.75×10^{-1}	186
SLC16A7	12	q14.1	7.33×10^{-1}	1606
INMT	7	p14.3	9.39×10^{-1}	541

4. Discussion

The use of admixed populations to detect variants that are absent or present at low frequency in Europeans populations, is widely documented (24–26). This is mainly due to particular linkage disequilibrium patterns found in admixed populations (27).

Several GWAS have identified nearly 60 susceptibility CRC loci among worldwide populations (28), but the single effect of these variants do not completely explain all the inherited variation that has been attributed to CRC. In the present study we were able to positively identify new candidate SNPs and genes showing evidence of association with CRC risk: a) one SNP associated with CRC, rs35797542, at genome-wide significance level ($p < 5 \times 10^{-8}$), corresponding to the Custom array. This variant has not been previously reported as associated with CRC in a GWAS and the nearest gene, *ZFAT*, is 25310 bp, and b) 16 other SNPs reached borderline statistical significance and 12 out of 16 overlapped codifying genes: rs16868695 with *SCUBE3*, rs255167 with *INMT*, rs73087778 with *FAM188B*, rs11860295 and rs3868142 with *PLEKHG4*, rs16957289 and rs16957304 with *KCTD19*, rs9922085, rs8047080, rs8052655 and rs7205526 with *LRRC36* and rs9922476 with *ATP6V0D1* (Table 1).

The rs35797542 variant has not been reported previously in association with any other disease. Its nearest gene is *ZFAT*, which is a Zinc Finger and AT-hook domain transcription factor. *ZFAT* locus, has been identified as correlated to ovarian cancer (29) and acute lymphocytic leukaemia (30), indicating an important role of *ZFAT* in cancer progression. Tsunoda and Shirasawa (31) reviewed the roles of *ZFAT* and concluded that it is an essential signalling molecule in haematopoietic, angiogenesis and cancer development.

Other nine of the associated SNPs lay in four genes located in cytoband 16q22.1 (*LRRC36*: rs7205526, rs8047080, rs8052655 and rs9922085; *PLEKHG4*: rs11860295 and rs3868142; *KCTD19*:

rs16957304 and rs16957289; *ATP6V0D1*: rs9922476) reached borderline statistical significance with CRC in our GWAS and those genes showed statistical significance in the SKAT analysis. This cytoband was reported as associated to CRC by Houlston et al. (32) in a meta-analysis of GWAS data of European samples, and replicated by Schmit et al. (33) in a GWAS of CRC in Hispanic samples including Mexicans.

Two out of four SNPs (rs8052655 and rs9922085) located in *LRRC36*, were previously reported in association with body-fat distribution in a large study of 344,369 individuals from five major ancestries and validation was performed in 132,177 European-ancestry individuals (34). This is why our result has to be taken with caution because this association might be due to an inbalance of obese individuals between cases and controls rather than a true association with CRC.

None of the SNPs in *PLEKHG4* (rs11860295 and rs3868142) or *KCTD19* (rs16957289 and rs16957304) nor the genes were previously reported in association through a GWAS with cancer or oncogenetic processes.

Although the SNP rs9922476 associated with *ATP6V0D1* gene was not reported previously, this gene was detected in a GWAS considering medication use as associated with the use of thyroid preparation (H03A) (35). This study was performed in a large sample of European ancestry from the UK Biobank. However, the same study showed a significant positive correlation between the use of the medication and higher values of BMI in addition to the fact that *ATP6V0D1* gene is located in the same cytoband as *LRRC36* mentioned before, so this association might be biased due to BMI. To discriminate the effect of the BMI further studies need to be done rather to confirm or refute the associations of these genes to CRC.

The 8 remaining SNPs are all non-coding variants (3 intronic, 2 upstream and 3 downs-

tream) and none of them were previously detected in a GWAS for CRC. The variant rs7589473 is located 18,834 bp upstream of gene *RBMS1*; this gene is a RNA binding motif single stranded interacting protein implied as a suppressor of colon cancer progression (36). Yu et al. (36) revealed that *RBMS1* is associated with poor prognosis in CRC patients because it is a positive regulator for multiple genes (including tumor suppressor AKAP12) involved in metastatic processes. The variant rs16868695 lies near *SCUBE3*, a gene showing up-regulation in a breast cancer subtype (37). Neither the SNPs rs255167 and rs73087778 nor their nearest genes (*INMT*, *FAM188B*) have been associated with any complex disease to date.

The rs1750424 variant is located close to *SLC10A2*, mutations in this genes have been found in a clinical trial of familial hypertriglyceridemia (38). This genetic disorder is typically associated with comorbidities such as hypertension, and obesity, conditions also present in CRC patients, therefore, this association has to be taken with caution. Also, other studies have observed the effect of mutations in *SLC10A2* over the risk of CRC. *SLC10A2* encodes an ileal sodium-dependent bile acid transporter which plays an important role in the absorption of bile acid in the large intestine, in vitro and in vivo experiments showed that malabsorption of bile acids and a loss of bile acids into the large intestine, increase cytotoxic in the colon and is strongly associated to colorectal adenomas, a precursor lesion of colorectal cancer (39,40). Besides, a bioinformatic analysis of clinical and genotypic data was performed by Wang et al. (41) detected an increased survival in CRC patients with certain prognostic signature, this signature includes alleles of 8 genes, *SLC10A2* among them. Lastly, the rs11173904 lies near *SLC16A7* gene which is part of the family of membrane transport proteins together with *SLC10A2*. Specifically *SLC16A7* has been detected in association to prostate cancer (42).

As stated above, only some of these SNPs and genes were previously reported as associated with cancer, stressing the lack of replication of associations studies among diverse

populations. This issue has been addressed by several authors. Sirugo et al. (43) attributes this lack of replication to population-specific variants, differences in linkage disequilibrium patterns as well as changes in allele frequency that arise as a product of genetic drift, local selection, or both. For the specific case of CRC, Lathroum et al. (44) have found heterogeneity due to ethnic diversity and concluded that genomic admixture has important implications in accurate diagnosis and treatment.

This study has strengths and limitations that should be mentioned. The main limitation is the sample size; in order to associations with lower effects (β <5), a larger sample is needed. The major strength is that the GWAS was conducted using an admixed Latin American population. To our knowledge it's the first GWAS of CRC conducted exclusively in Mexican samples.

5. Conclusions

To our knowledge it is the first GWAS of CRC conducted exclusively in Mexican samples. We found novel candidate variants and their associated genes which should be further investigated to confirm their role in CRC aetiology. The findings are relevant because one new SNP (rs35797542) and its nearest gene, *ZFAT*, were associated with CRC in Mexicans. This gene was previously associated with cancer and other complex phenotypes. The other 4 genes reported as associated (*LRRC36*, *PLEKHG4*, *KCTD19* and *ATP6V0D1*) showed a wide variability of supporting evidence to be related to cancer and the shared cytoband (16q22.1) is a highly attractive opportunity to continue studying it. Some of the associated loci play important roles in cell proliferation, differentiation, development and oncogenesis, but for many of them, functional assay will be determinant to define their role in the etiology and development of CRC.

Our study contributes to fully comprehend the genetic causes of CRC. It also highlights the importance of conducting studies in diverse worldwide populations because of the importance of ethnic-specific genomic variation undetected in association with CRC in non-admixed samples. Validation of these SNPs and genes in other admixed Latin American populations would be important to add evidence of the pool of variants associated with CRC in this study.

Declarations

Authors Contributions: Conceptualization, M.S. and A.R.M.; Data Curation, VC, IQ and RC; Formal Analysis, V.C. and R.C.; Funding Acquisition; A.C., M.S. and A.R.M.; Investigation I.Q. and C.R.; Methodology, V.C. and R.C.; Project Administration, A.C., M.S. and A.R.M.; Resources, I.Q., C.R., P.L.P., I.S.G.G, C.M.M., R.O.L., Y.J.R., P.R.F., E.C.M., J.F.G.G, S.C.C., N.K.M.S., J.H.S.C and F.M.AM.; Visualization, V.C.;Writing original draft,

V.C. and M.S.; Writing review & editing R.C., A.C. and A.R.M.

Data availability statement: The datasets generated and analyzed during the current study are available from corresponding authors on reasonable request.

Funding: This research was funded by the Seventh Framework Programme of the European Commission. CHIBCHA project number: 223 678. Genetic study of Common Hereditary Bowel Cancers in Hispania and the Americas.

Acknowledgments: Members of the CHIBCHA Consortium: Ian Tomlinson (University of Edinburgh, UK); Luis Carvalal-Carmona (University of California, Davis, USA), Ma. Magdalena Echeverry de Polanco, Mabel Elena Bohórquez, Rodrigo Prieto, Angel Criollo, Carolina Ramírez, Ana Patricia Estrada, Jhon Jairo Suárez (Universidad del Tolima, Colombia); Augusto Rojas Martinez, Rocío Ortiz López (Tecnológico de Monterrey, Mexico); Silvia Rogatto, Samuel Aguiar Jnr, Ericka Maria Monteiro Santos (São Paulo State University, Botucatu, Brazil); Monica Sans, Valentina Colistro, Pedro C. Hidalgo, Patricia Mut (University of the Republic, Uruguay); Angel Carracedo, Clara Ruiz Ponte, Ines Quntela Garcia (-University of Santiago de Compostela, Spain); Sergi Castellvi-Bel (University of Barcelona, Barcelona, Catalonia, Spain); Manuel Teixeira (Portuguese Oncology Institute, Portugal).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Global Cancer Observatory. (2019) Available at: https://gco.iarc.fr/ [Accessed December 2, 2020]

2. Sánchez-Barriga JJ. Mortality trends and risk of dying from colorectal cancer in the seven socioeconomic regions of Mexico, 2000-2012. Rev Gastroenterol Mex (2017) 82:217–225. doi:10.1016/j.rgmx.2016.10.005

3. GWAS Catalog. (2020) Available at: https://www.ebi.ac.uk/gwas/ [Accessed April 4, 2020]

4. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet (2008) 40:623–630. doi:10.1038/ng.111

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature (2009) 461:747–753. doi:10.1038/nature08494

6. Peterson AC, Di Rienzo A, Lehesjokl AE, De La Chaelle A, Slatkin M, Frelmer NB. The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet (1995) 4:887–894. doi:10.1093/hmg/4.5.887

7. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet (2006) 38:556–560. doi:10.1038/ng1770

8. Stumpf MPHH, Goldstein DB. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol (2003) 13:1–8. doi:10.1016/S0960-9822(02)01404-5

9. Goddard KAB, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms five populations. Am J Hum Genet (2000) 66:216–234. doi:10.1086/302727

10. Ko A, Cantor RM, Weissglas-Volkov D, Nikkola E, Reddy PMVL, Sinsheimer JS, Pasaniuc B, Brown R, Alvarez M, Rodriguez A, et al. Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Commun (2014) 5:3983. doi:10.1038/ncomms4983

11. Weissglas-Volkov D, Aguilar-Salinas CA, Nikkola E, Deere KA, Cruz-Bautista I, Arellano-Campos O, Muñoz-Hernandez LL, Gomez-Munguia L, Ordoñez-Sánchez ML, Linga Reddy PMV, et al. Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. J Med Genet (2013) 50:298–308. doi:10.1136/jmedgenet-2012-101461

12. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris P, Zondervan KT, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc (2010) 5:1564–1573. doi:10.1038/nprot.2010.116.Data

13. Chang CC, Chow CC, Tellier LCAMC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience (2015) 4:7.doi:10.1186/s13742-015-0047-8

14. Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Debakker P, Daly MJ, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet (2007) 81:559–575. doi:10.1086/519795

15. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet (2012) 91:122–138. doi:10.1016/j.ajhg.2012.05.024

16. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet (2006) 2:e190. doi:10.1371/journal.pgen.0020190

17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet (2006) 38:904–909. doi:10.1038/ng1847

18. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet (2007) 80:1171–1178. doi:10.1086/518564

19. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet (2007) 81:1084–1097. doi:10.1086/521987

20. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. Nat Genet (2016) 48:1284–1287. doi:10.1038/ng.3656

21. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet (2013) 92:841–853. doi:10.1016/j.ajhg.2013.04.015

22. Colistro V, Rojas-Martínez A, Carracedo A, Tomlinson I, Carvajal-Carmona L, Cruz R, Sans M. Population structure and relatedness estimates in a Mexican sample. Ann Hum Genet (2021) doi:10.1111/ahg.12421

23. Ensembl genome browser 99. (2019) Available at: https://www.ensembl.org/index.html [Accessed November 4, 2020]

24. Pasaniuc B, Zaitlen N, Lettre G, Chen GK, Tandon A, Linda WH, Ruczinski I, Fornage M, Siscovick DS, Zhu X, et al. Enhanced Statistical Tests for GWAS in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet (2011) 7: doi:10.1371/journal.pgen.1001371

25. Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, Huntsman S, Nishimura KK, Gourraud PA, Mohajeri K, et al. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos. J Allergy Clin Immunol (2015) 135:1502–1510. doi:10.1016/j.jaci.2014.10.033

26. Pulit SL, Voight BF, de Bakker PIW. Multiethnic Genetic Association Studies Improve Power for Locus Discovery. PLoS One (2010) 5:e12600. doi:10.1371/journal.pone.0012600

27. Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet (2001) 68:198–207. doi:10.1086/316935

28. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, Farrington S, Svinti V, Palles C, Orlando G, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun (2019) 10:1–15. doi:10.1038/s41467-019-09775-w

29. Ramakrishna M, Williams LH, Boyle SE, Bearfoot JL, Sridhar A, Speed TP, Gorringe KL, Campbell IG. Identification of Candidate Growth Promoting Genes in Ovarian Cancer through Integrated Copy Number and Expression Analysis. PLoS One (2010) 5:e9983. doi:10.1371/journal.pone.0009983

30. Fujimoto T, Doi K, Koyanagi M, Tsunoda T, Takashima Y, Yoshida Y, Sasazuki T, Shirasawa S. ZFAT is an antiapoptotic molecule and critical for cell survival in MOLT-4 cells. FEBS Lett (2009) 583:568–572. doi:10.1016/j.febslet.2008.12.063

31. Tsunoda T, Shirasawa S. Roles of ZFAT in haematopoiesis, angiogenesis and cancer development. Anticancer Res (2013) 33:2833–2838.

32. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S, Chandler I, Vijayakrishnan J, Sullivan K, Penegar S, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet (2008) 40:1426–1435.doi:10.1038/ng.262

33. Schmit SL, Schumacher FR, Edlund CK, Conti D V., Ihenacho U, Wan P, Van Den Berg D, Casey G, Fortini BK, Lenz H-JJ, et al. Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis (2016) 37:547–556. doi:10.1093/carcin/bgw046

34. Justice AE, Karaderi T, Highland HM, Young KL, Graff M, Lu Y, Turcot V, Auer PL, Fine RS, Guo X, et al. Protein-coding variants implicate novel genes related to lipid homeostasis contributing to body-fat distribution. Nat Genet (2019) 51:452–469. doi:10.1038/s41588-018-0334-2

35. Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, Yang J, Visscher PM, Wray NR. Genome-wide association study of medication-use and associated disease in the UK Biobank. Nat Commun (2019) 10:1–10. doi:10.1038/s41467-019-09572-5

36. Yu J, Navickas A, Asgharian H, Culbertson B, Fish L, Garcia K, Olegario JP, Dermit M, Dodel M, Hänisch B, et al. Rbms1 suppresses colon cancer metastasis through targeted stabilization of its mRNA regulon. Cancer Discov (2020) 10:1410–1423. doi:10.1158/2159-8290.CD-19-1375

37. Gao C, Zhuang J, Li H, Liu C, Zhou C, Liu L, Feng F, Sun C, Wu J. Development of a risk scoring system for evaluating the prognosis of patients with Her2-positive breast cancer. Cancer Cell Int (2020) 20:121.doi:10.1186/s12935-020-01175-1

38. Love MW, Craddock AL, Angelin B, Brunzell JD, Duane WC, Dawson PA. Analysis of the ileal bile acid transporter gene, SLC10A2, in subjects with familial hypertriglyceridemia. Arterioscler Thromb Vasc Biol (2001) 21:2039–2045. doi:10.1161/hq1201.100262

39. Wang W, Xue S, Ingles SA, Chen Q, Diep AT, Frankl HD, Haile RW, Stolz A. An association between genetic polymorphisms in the ileal sodium-dependent bile acid transporter gene and the risk of colorectal adenomas. Cancer Epidemiol Biomarkers Prev (2001) 10:931–936.

40. Raufman J-P, Dawson PA, Rao A, Drachenberg CB, Heath J, Shang AC, Hu S, Zhan M, Polli JE, Cheng K. Slc10a2-null mice uncover colon cancer-promoting actions of endogenous fecal bile acids. Carcinogenesis (2015) 36:1193–1200.doi:10.1093/carcin/bgv107

41. Wang J, Yu S, Chen G, Kang M, Jin X, Huang Y, Lin L, Wu D, Wang L, Chen J. A novel prognostic signature of immune-related genes for patients with colorectal cancer. J Cell Mol Med (2020) 24:8491–8504. doi:10.1111/jcmm.15443

42. Pértega-Gomes N, Vizcaino JR, Felisbino S, Warren AY, Shaw G, Kay J, Whitaker H, Lynch AG, Fryer L, Neal DE, et al. Epigenetic and oncogenic regulation of SLC16A7 (MCT2) results in protein over-expression, impacting on signalling and cellular phenotypes in prostate cancer. Oncotarget (2015) 6:21675–21684. doi:10.18632/oncotarget.4328

43. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell (2019) 177:26–31. doi:10.1016/j.cell.2019.02.048

44. Lathroum L, Ramos-Mercado F, Hernandez-Marrero J, Villafaña M, Cruz-Correa M. Ethnic and Sex Disparities in Colorectal Neoplasia Among Hispanic Patients Undergoing Screening Colonoscopy. Clin Gastroenterol Hepatol (2012) 10:997–1001. doi:10.1016/j.cgh.2012.04.015

Supplementary material.

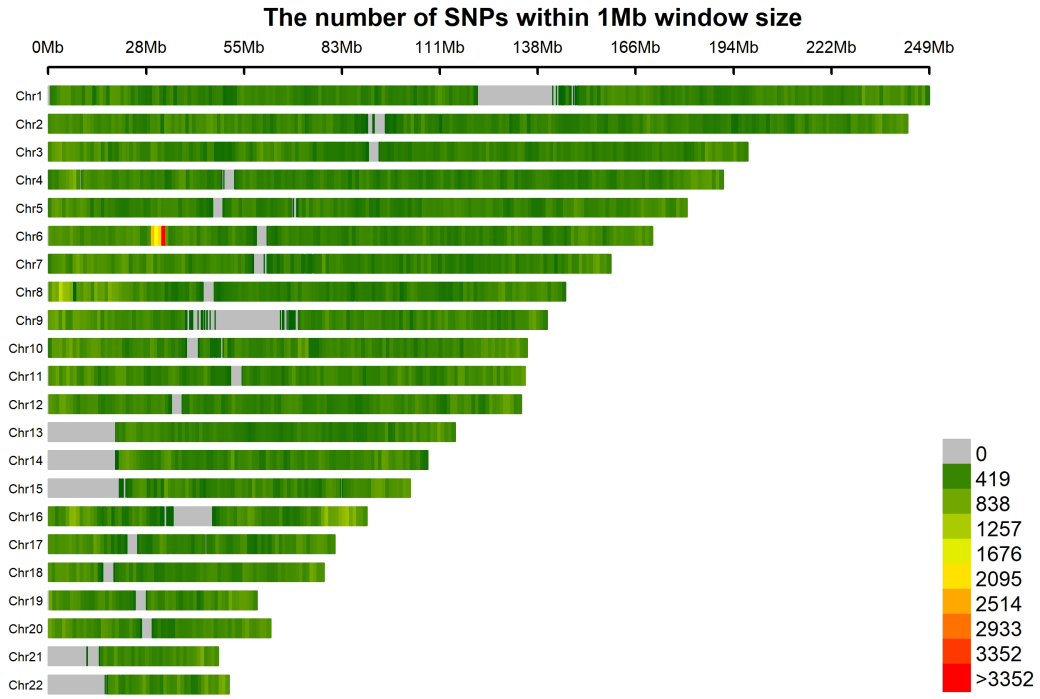

Figure S1. Density of variants distribution along the 22 autosomal chromosomes. SNP from both arrays are considered jointly.

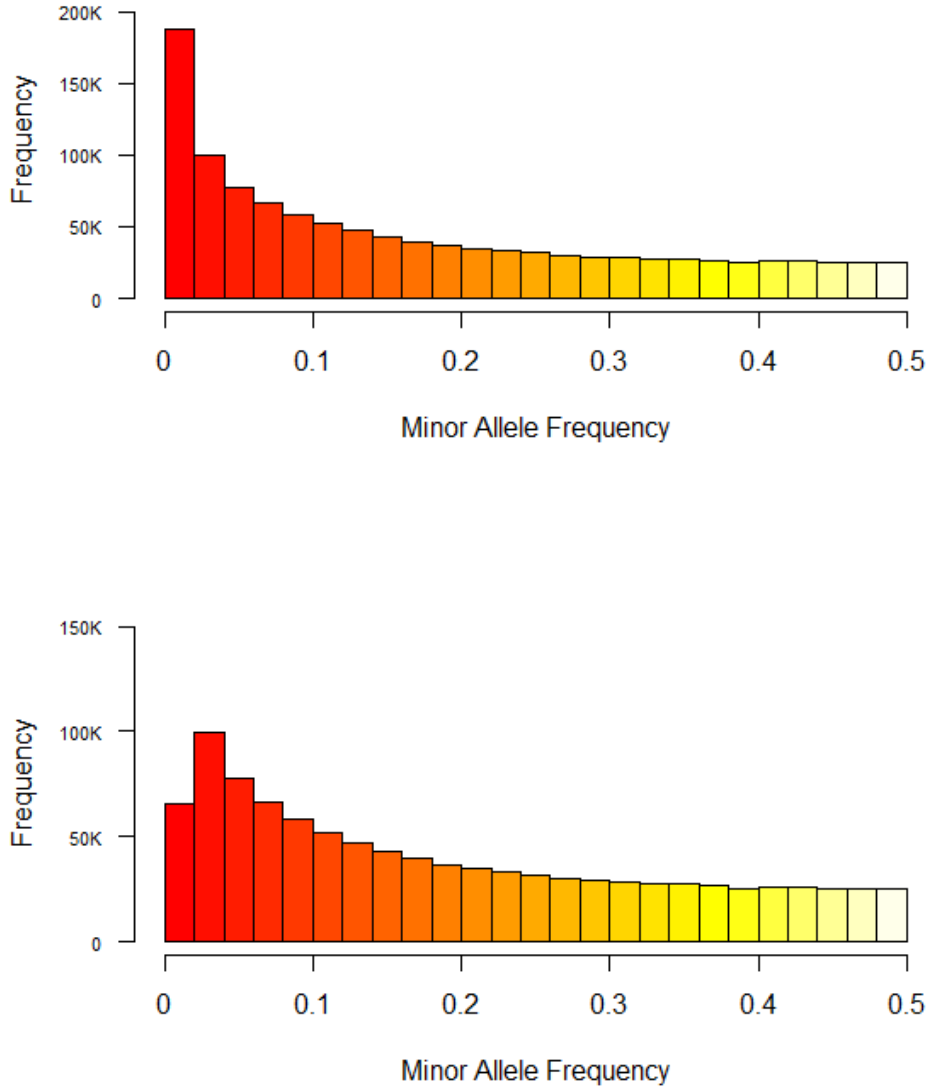

Figure S2. Histograms of MAF distribution of SNPs. All SNPs (upper histogram), filtering out SNPs with MAF<0.01 (lower histogram).

Capítulo 5

Conclusiones y perspectivas

En primer lugar queremos destacar que los resultados de todos los capítulos de esta tesis reafirman las ventajas de incluir poblaciones mestizadas en estudios de asociación genética; también se evidencia la esencialidad de considerar la historia demográfica de la población para poder realizar estudios de genética poblacional.

Por un lado, mostramos que a nivel poblacional los valores de ancestría detectados en las muestras mexicanas son coherentes con la historia de la población de México; al realizar un análisis de ancestría por regiones genómicas candidatas a CRC, no hay un patrón común, sino que cada una presenta proporciones ancestrales diferentes. Esto se repitió también con las estimaciones de heterocigosidad y en el largo y la cantidad de los bloques haplotípicos. Todo esto evidencia la necesidad de incorporar estimaciones de la ancestría individual local (por individuo y por locus) en los estudios poblacionales, así como también la importancia de interpretar los resultados a la luz de la estructura genética poblacional ya que de no ser así podríamos detectar regiones como asociadas pero no por su importancia funcional en la

patología, sino por su presencia diferencial entre individuos con ancestrías diferentes en la región genómica en cuestión.

En linea con esta conclusión, también mostramos la importancia de considerar la mezcla poblacional en los análisis genético poblacionales ya que como quedó demostrado en las estimaciones de parentesco, las mismas son fuertemente influenciadas por la ancestría genética en poblaciones mestizadas. En particular, observamos que una mayor proporción de ancestría indígena era confundido con un mayor grado de consanguinidad por los *software* estándar, y esto se corrige al aplicar *software* específicamente diseñados para trabajar con poblaciones mestizadas.

Por último, este trabajo incluye el primer GWAS de CRC realizado exclusivamente en población Mexicana, donde detectamos un SNP y su gen asociado como candidatos a ser responsables de la presentación y desarrollo del CRC (rs35797542 y *ZFAT*, en 8q24.22). Además, postulamos otros 4 genes (*LRRC36, ATP6V0D1,KCTD19* y *PLEKHG4*) como posibles candidatos, los cuales sería necesario analizar en profundidad para poder determinar cuales mutaciones en esos genes tienen un impacto en el CRC. Entre los resultados del GWAS se destaca por un lado el potencial de las poblaciones mestizadas en estudios de asociación ya que permiten detectar variantes no detectadas en poblaciones no mestizadas; y por otro la ausencia de reproducibilidad de los resultados reportados en poblaciones no mestizadas. Esto último es de suma importancia ya que reafirma que los procedimientos metodológicos al lidiar con poblaciones mestizadas son particulares y específicos para poblaciones con esta condición, diferenciándose de aquellos procedimientos para poblaciones no mestizadas.

Además, este trabajo contribuye a la realización de análisis de asociación con individuos multiétnicos; estos estudios arrojan nuevos resultados que sumados a los ya existentes podrían explicar un mayor porcentaje de la heredeabilidad del CRC. En este sentido, sería de esperar

que se realicen en el futuro más estudios de asociación con individuos de diversas ancestrías y/o estudios de metanálisis que sistematicen los datos publicados previamente y los analice a la luz de la diversidad étnica de los individuos.

Todas estas conclusiones permiten pensar nuevos caminos a seguir para poder avanzar en la comprensión de los procesos genéticos que condicionan al CRC. Por ejemplo, se podría estimar la ancestría local individual a lo largo de todo el genoma y con ello generar un mapa de recombinación específico de la población Mexicana. El mismo sería un insumo valioso al momento de mapear genes o de determinar patrones de ligamiento en el genoma de individuos mexicanos.

Capítulo 6

Bibliografía[1]

Ahlbom, A., Lichtenstein, P., Malmstrom, H., Feychting, M., Pedersen, N. L., & Hemminki, K. (1997). Cancer in Twins: Genetic and Nongenetic Familial Risk Factors. JNCI Journal of the National Cancer Institute, 89(4), 287–293. https://doi.org/10.1093/jnci/89.4.287

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, P., Zondervan, K. T., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. Nature Protocols, 5(9), 1564–1573. https://doi.org/10.1038/nprot.2010.116.Data

Bertoni, B., Jin, L. I., Chakraborty, R., & Sans, M. (2005). Directional mating and a rapid male population expansion in a hybrid Uruguayan population. American Journal of Human Biology: The Official Journal of the Human Biology Council, 17(6), 801–808. https://doi.org/10.1002/ajhb.20443

Bonilla, C., Bertoni, B., González, S., Cardoso, H., Brum-Zorrilla, N., & Sans, M. (2004). Substancial native american female contribution to the population of tacuarembó, Uruguay, reveals past episodes of sex-biased gene flow. American Journal of Human Biology, 16, 289–297. https://doi.org/10.1002/ajhb.20025

[1]La bibliografía citada únicamente en los artículos no está incluida en este capítulo.

Bonilla, C., Bertoni, B., Hidalgo, P. C., Artagaveytia, N., Ackermann, E., Barreto, I., Cancela, P., Cappetta, M., Egaña, A., Figueiro, G., Heinzen, S., Hooker, S., Román, E., Sans, M., & Kittles, R. A. (2015). Breast cancer risk and genetic ancestry: A case-control study in Uruguay. BMC Women's Health, 15(1). https://doi.org/10.1186/s12905-015-0171-8

Botteri, E., Iodice, S., Raimondi, S., Maisonneuve, P., & Lowenfels, A. B. (2008). Cigarette Smoking and Adenomatous Polyps: A Meta-analysis. Gastroenterology, 134(2), 388-395.e3. https://doi.org/10.1053/j.gastro.2007.11.007

Boyle, P., & Leon, M. E. (2002). Epidemiology of colorectal cancer. In British Medical Bulletin (Vol. 64, pp. 1–25). Br Med Bull. https://doi.org/10.1093/bmb/64.1.1

Broderick, P., Carvajal-Carmona, L., Pittman, A. M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S., Jaeger, E., Vijayakrishnan, J., Kemp, Z., Gorman, M., Chandler, I., Papaemmanuil, E., Penegar, S., Wood, W., Sellick, G., … Houlston, R. S. (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nature Genetics, 39(11), 1315–1317. https://doi.org/10.1038/ng.2007.18

Broeke, S. W. T., Klift, H. M. V., Tops, C. M. J., Aretz, S., Bernstein, I., Buchanan, D. D., Chapelle, A. Dela, Capella, G., Clendenning, M., Engel, C., Gallinger, S., Garcia, E. G., Figueiredo, J. C., Haile, R., Hampel, H. L., Hopper, J. L., Hoogerbrugge, N., Doeberitz, M. V. K., Marchand, L. Le, … Win, A. K. (2018). Cancer Risks for PMS2-associated lynch syndrom. Journal of Clinical Oncology, 36(29), 2961–2968. https://doi.org/10.1200/JCO.2018.78.4777

Bryc, K., Velez, C., Karafet, T., Moreno-estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D., & Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences, 107, 8954–8961. https://doi.org/10.1073/pnas.0914618107

Carlson, M. O., Montilla-Bascon, G., Hoekenga, O. A., Tinker, N. A., Poland, J., Baseggio, M., Sorrells, M. E., Jannink, J. L., Gore, M. A., & Yeats, T. H. (2019). Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (Avena sativa L.). G3: Genes, Genomes, Genetics, 9(9), 2963–2975. https://doi.org/10.1534/g3.119.400228

Carvajal-Carmona, L. G., Cazier, J.-B., Jones, A. M., Howarth, K., Broderick, P., Pittman, A., Dobbins, S., Tenesa, A., Farrington, S., Prendergast, J., Theodoratou, E., Barnetson, R., Conti, D., Newcomb, P. A., Hopper, J. L., Jenkins, M. A., Gallinger, S., Duggan, D. J., Campbell, H., ... Tomlinson, I. (2011). Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Human Molecular Genetics, 20(14), 2879–2888. https://doi.org/10.1093/hmg/ddr190

Cesarani, A., Gaspa, G., Pauciullo, A., Degano, L., Vicario, D., & Macciotta, N. P. P. (2020). Genome-wide analysis of homozygosity regions in european simmental bulls. Journal of Animal Breeding and Genetics, 138(1). https://doi.org/10.1111/jbg.12502

Chakraborty, R., & Weiss, K. M. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proceedings of the National Academy of Sciences of the United States of America, 85(23), 9119–9123. https://doi.org/10.1073/pnas.85.23.9119

Chan, D. S. M., Lau, R., Aune, D., Vieira, R., Greenwood, D. C., Kampman, E., & Norat, T. (2011). Red and Processed Meat and Colorectal Cancer Incidence: Meta-Analysis of Prospective Studies. PLoS ONE, 6(6), e20456. https://doi.org/10.1371/journal.pone.0020456

Chang, C. C., Chow, C. C., Tellier, L. C. A. M. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Colistro, V., Mut, P., Hidalgo, P. C., Carracedo, A., Quintela, I., Rojas-Martínez, A., & Sans, M. (2020). Differential admixture in Latin American populations and its impact on the study of colorectal cancer. Genetics and Molecular Biology, 43(4), 1–9. https://doi.org/10.1590/1678-4685-GMB-2020-0143

Colistro, V., Rojas-Martínez, A., Carracedo, A., CHIBCHA Consortium, Tomlinson, I., Carvajal-Carmona, L., Cruz, R., & Sans, M. (2021). Population structure and relatedness estimates in a Mexican sample. Ann Hum Genet.;1–4. https://doi.org/10.1111/ahg.12421

CONAPO. (2009). Encuesta Nacional de la Dinámica Demográfica 2009 Panorama sociodemográfico de México. Instituto Nacional de Estadística y Geografía. http://internet.contenidos.inegi.org.mx/contenidos/Productos/prod_serv/contenidos/espanol/bvinegi/productos/encuestas/hogares/enadid/enadid2009/702825495602.pdf

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., ... Donnelly, P. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464(7289), 713–720. https://doi.org/10.1038/nature08979

Cui, R., Okada, Y., Jang, S. G., Ku, J. L., Park, J. G., Kamatani, Y., Hosono, N., Tsunoda, T., Kumar, V., Tanikawa, C., Kamatani, N., Yamada, R., Kubo, M., Nakamura, Y., & Matsuda, K. (2011). Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut, 60(6), 799–805. https://doi.org/10.1136/gut.2010.215947

Da Luz, J., Kimura, E. M., Costa, F. F., Sonati, M. F., & Sans, M. (2010). Beta-globin gene cluster haplotypes in Afro-Uruguayans from two geographical regions (South and North). American Journal of Human Biology: The Official Journal of the Human Biology Council, 22(1), 124–128. https://doi.org/10.1002/ajhb.20961

Data Visualizations | Institute for Health Metrics and Evaluation. (2020). http://www.healthdata.org/results/data-visualizations

Dowty, J. G., Win, A. K., Buchanan, D. D., Lindor, N. M., Macrae, F. A., Clendenning, M., Antill, Y. C., Thibodeau, S. N., Casey, G., Gallinger, S., Marchand, L. Le, Newcomb, P. A., Haile, R. W., Young, G. P., James, P. A., Giles, G. G., Gunawardena, S. R., Leggett, B. A., Gattas, M., ... Jenkins, M. A. (2013). Cancer Risks for MLH1 and MSH2 Mutation Carriers. Human Mutation, 34(3), 490–497. https://doi.org/10.1002/humu.22262

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews. Genetics, 11(6), 446–450. https://doi.org/10.1038/nrg2809

Escalante Gonzalbo, P., García Martínez, B., Jáuregui, L., Zoraida Vázquez, J., Speckman Guerra, E., Garciadiego, J., & Aboites Aguilar, L. (2004). Nueva Historia Mínima de México (Primera Ed). El Colégio de México A.C. Centro de Estudios Históricos.

Farreas-Rozman. (2013). Medicina Interna. Metabolismo y Nutrición. Endocrinología. (Ciril Rozman Borstnar & Francesc Cardellach (Eds.); 17th Editi). Elsevier.

Fejerman, L., Carnese, F. R., Goicoechea, A. S., Avena, S. a, Dejean, C. B., & Ward, R. H. (2005). African ancestry of the population of Buenos Aires. American Journal of Physical Anthropology, 128(1), 164–170. https://doi.org/10.1002/ajpa.20083

Fernandez-Rozadilla, C., Cazier, J.-B., Tomlinson, I. P., Carvajal-Carmona, L. G., Palles, C., Lamas, M. J., Baiget, M., López-Fernández, L. a, Brea-Fernández, A., Abulí, A., Bujanda, L., Clofent, J., Gonzalez, D., Xicola, R., Andreu, M., Bessa, X., Jover, R., Llor, X., Moreno, V., ... Ruiz-Ponte, C. (2013). A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. BMC Genomics, 14, 55. https://doi.org/10.1186/1471-2164-14-55

Flossmann, E., & Rothwell, P. M. (2007). Effect of aspirin on long-term risk of colorectal cancer: consistent evidence from randomised and observational studies. Lancet, 369(9573), 1603–1613. https://doi.org/10.1016/S0140-6736(07)60747-8

Fodde, R. (2002). The APC gene in colorectal cancer. European Journal of Cancer, 38(7), 867–871. https://doi.org/10.1016/S0959-8049(02)00040-0

Fransén, K., Klintenäs, M., Österström, A., Dimberg, J., Monstein, H. J., & Söderkvist, P. (2004). Mutation analysis of the BRAF, ARAF and RAF-1 genes in human colorectal adenocarcinomas. Carcinogenesis, 25(4), 527–533. https://doi.org/10.1093/carcin/bgh049

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. a, Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., & Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. Nature Genetics, 36(4), 388–393. https://doi.org/10.1038/ng1333

Gascue, C., Mimbacas, A., Sans, M., Gallino, J. P., Bertoni, B., Hidalgo, P., Cardoso, H., & Galling, J. P. (2005). Frequencies of the four major Amerindian mtDNA haplogroups in the population of Montevideo, Uruguay. Human Biology, 77(6), 873–878.https://doi.org/10.1353/hub.2006.0015

Global Cancer Observatory. (2019). https://gco.iarc.fr/

Goicoechea, A. S., Carnese, F. R., Dejean, C., Avena, S. a, Weimer, T. a, Franco, M. H., Callegari-Jacques, S. M., Estalote, a C., Simões, M. L., Palatnik, M., & Salzano, F. M. (2001). Genetic relationships between Amerindian populations of Argentina. American Journal of Physical Anthropology, 115(2), 133–143. https://doi.org/10.1002/ajpa.1063

Gorostiza, A., Acunha-Alonzo, V., Regalado-Liu, L., Tirado, S., Granados, J., Sámano, D., Rangel-Villalobos, H., & González-Martín, A. (2012). Reconstructing the History of Mesoamerican Populations through the Study of the Mitochondrial DNA Control Region. PLoS ONE, 7(9), e44666. https://doi.org/10.1371/journal.pone.0044666

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. a, & Bustamante, C. D. (2011). Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America, 108(29), 11983–11988. https://doi.org/10.1073/pnas.1019276108

GWAS Catalog. (2020). https://www.ebi.ac.uk/gwas/docs/diagram-downloads

Haiman, C. A., Le Marchand, L., Yamamato, J., Stram, D. O., Sheng, X., Kolonel, L. N., Wu, A. H., Reich, D., & Henderson, B. E. (2007). A common genetic risk factor for colorectal and prostate cancer. Nature Genetics, 39(8), 954–956. https://doi.org/10.1038/ng2098

Hannan, L. M., Jacobs, E. J., & Thun, M. J. (2009). The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. Cancer Epidemiology Biomarkers and Prevention, 18(12), 3362–3367. https://doi.org/10.1158/1055-9965.EPI-09-0661

Henrikson, N. B., Webber, E. M., Goddard, K. A., Scrol, A., Piper, M., Williams, M. S., Zallen, D. T., Calonge, N., Ganiats, T. G., Janssens, A. C. J. W., Zauber, A., Lansdorp-Vogelaar, I., Van Ballegooijen, M., & Whitlock, E. P. (2015). Family history and the natural history of colorectal cancer: Systematic review. In Genetics in Medicine (Vol. 17, Issue 9, pp. 702–712). Nature Publishing Group. https://doi.org/10.1038/gim.2014.188

Hidalgo, P. C., Bengochea, M., Abilleira, D., Cabrera, A., & Alvarez, I. (2005). Genetic Admixture Estimate in the Uruguayan Population Based on the Loci LDLR , GYPA , HBGG , GC and D7S8. International Journal of Human Genetics, 5(3), 217–222. https://doi.org/10.1080/09723757.2005.11885929

Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., Carvajal-Carmona, L., Howarth, K., Jaeger, E., Spain, S. L., Walther, A., Barclay, E., Martin, L., Gorman, M., Domingo, E., ... Dunlop, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nature Genetics, 40(12), 1426–1435. https://doi.org/10.1038/ng.262

Huang, C.-T., Esvelt Klos, K., & Huang, Y.-F. (2020). Genome-Wide Association Study Reveals the Genetic Architecture of Seed Vigor in Oats. G3; Genes|Genomes|Genetics, 10(12), g3.401602.2020. https://doi.org/10.1534/g3.120.401602

Hur, S. J., Yoon, Y., Jo, C., Jeong, J. Y., & Lee, K. T. (2019). Effect of Dietary Red Meat on Colorectal Cancer Risk—A Review. Comprehensive Reviews in Food Science and Food Safety, 18(6), 1812–1824. https://doi.org/10.1111/1541-4337.12501

International HapMap Project. (2020). https://www.genome.gov/10001688/international-hapmap-project

Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Walther, A., Spain, S., Pittman, A., Kemp, Z., Sullivan, K., Heinimann, K., Lubbe, S., Domingo, E., Barclay, E., Martin, L., Gorman, M., Chandler, I., Vijayakrishnan, J., … Tomlinson, I. (2008). Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nature Genetics, 40(1), 26–28. https://doi.org/10.1038/ng.2007.41

Jia, W. H., Zhang, B., Matsuo, K., Shin, A., Xiang, Y. B., Jee, S. H., Kim, D. H., Ren, Z., Cai, Q., Long, J., Shi, J., Wen, W., Yang, G., Delahanty, R. J., Ji, B. T., Pan, Z. Z., Matsuda, F., Gao, Y. T., Oh, J. H., … Zheng, W. (2013). Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. Nature Genetics, 45(2), 191–196. https://doi.org/10.1038/ng.2505

Khlestkin, V. K., Erst, T. V., Rozanova, I. V., Efimov, V. M., & Khlestkina, E. K. (2020). Genetic loci determining potato starch yield and granule morphology revealed by genome-wide association study (GWAS). PeerJ, 8. https://doi.org/10.7717/peerj.10286

Klarskov, L., Holck, S., Bernstein, I., Okkels, H., Rambech, E., Baldetorp, B., & Nilbert, M. (2011). Challenges in the identification of MSH6-associated colorectal cancer: Rectal location, less typical histology, and a subset with retained mismatch repair function. American Journal of Surgical Pathology, 35(9), 1391–1399. https://doi.org/10.1097/PAS.0b013e318225c3f0

Kolonel, L. N., Henderson, B. E., Hankin, J. H., Nomura, A. M. Y., Wilkens, L. R., Pike, M. C., Stram, D. O., Monroe, K. R., Earle, M. E., & Nagamine, F. S. (2000). A multiethnic cohort in Hawaii and Los Angeles: Baseline characteristics. American Journal of Epidemiology, 151(4), 346–357. https://doi.org/10.1093/oxfordjournals.aje.a010213

Kudinov, A. A., Dementieva, N. V., Mitrofanova, O. V., Stanishevskaya, O. I., Fedorova, E. S., Larkina, T. A., Mishina, A. I., Plemyashov, K. V., Griffin, D. K., & Romanov, M. N. (2019). Genome-wide association studies targeting the yield of extraembryonic fluid and production traits in Russian White chickens. BMC Genomics, 20(1). https://doi.org/10.1186/s12864-019-5605-5

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., Dash, C., Giardiello, F. M., Glick, S., Johnson, D., Johnson, C. D., Levin, T. R., Pickhardt, P. J., Rex, D. K., Smith, R. A., Thorson, A., & Winawer, S. J. (2008). Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline From the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology, 134(5), 1570–1595. https://doi.org/10.1053/j.gastro.2008.02.002

Li, F. P. (1995). Phenotypes, Genotypes, and Interventions for Hereditary Cancers. Cancer Epidemiology and Prevention Biomarkers, 4(6), 579–582.https://cebp.aacrjournals.org/content/cebp/4/6/579.full.pdf

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., & Hemminki, K. (2000). Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. New England Journal of Medicine, 343(2), 78–85. https://doi.org/10.1056/nejm200007133430201

Lisker, R., Ramírez, E., & Babinsky, V. (1996). Genetic Structure of Autochthonous Populations of Meso-America: Mexico. Human Biology, 68(3), 395–404.http://www.jstor.org/stable/41465484

Liu, Z., Yang, N., Yan, Y., Li, G., Liu, A., Wu, G., & Sun, C. (2019). Genome-wide association analysis of egg production performance in chickens across the whole laying period. BMC Genetics, 20(1). https://doi.org/10.1186/s12863-019-0771-7

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753. https://doi.org/10.1038/nature08494

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., McKeigue, P. M., Shriver, M. D., & Parra, E. J. (2007). A genomewide admixture mapping panel for Hispanic/Latino populations. American Journal of Human Genetics, 80(6), 1171–1178. https://doi.org/10.1086/518564

Martínez-González, L. J., Saiz, M., Alvarez-Cubero, M. J., Gómez-Martín, A., Alvarez, J. C., Martínez-Labarga, C., & Lorente, J. A. (2012). Distribution of Y chromosomal STRs loci in Mayan and Mestizo populations from Guatemala. Forensic Science International. Genetics, 6(1), 136–142. https://doi.org/10.1016/j.fsigen.2011.04.003

Martínez, M. E. (2005). Primary prevention of colorectal cancer: lifestyle, nutrition, exercise. In Recent results in cancer research. (Vol. 166). Recent Results Cancer Res. https://doi.org/10.1007/3-540-26980-0_13

Martínez Marignac, V. L., Bertoni, B., Parra, E. J., & Bianchi, N. O. (2004). Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. Human Biology, 76(4), 543–557. https://doi.org/10.1353/hub.2004.0058

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nature Genetics, 44(3), 243–246. https://doi.org/10.1038/ng.1074.Differential

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. In Nature Reviews Genetics (Vol. 9, Issue 5, pp. 356–369). Nat Rev Genet. https://doi.org/10.1038/nrg2344

McKeigue, P. M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. American Journal of Human Genetics, 63(1), 241–251. https://doi.org/10.1086/301908

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science, 344(6189), 1280–1285. https://doi.org/10.1126/science.1251688

Mörner, M. (1967). Race mixture in the history of Latin América (Little, Br).

Norris, E. T., Wang, L., Conley, A. B., Rishishwar, L., Mariño-Ramírez, L., Valderrama-Aguirre, A., & Jordan, I. K. (2018). Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics, 19(Suppl 8). https://doi.org/10.1186/s12864-018-5195-7

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. a, Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J. M. J., & Reich, D. (2004). Methods for high-density admixture mapping of disease genes. American Journal of Human Genetics, 74(5), 979–1000. https://doi.org/10.1086/420871

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. PLoS Genetics, 2(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Peters, U., Jiao, S., Schumacher, F. R., Hutter, C. M., Aragaki, A. K., Baron, J. A., Berndt, S. I., Bézieau, S., Brenner, H., Butterbach, K., Caan, B. J., Campbell, P. T., Carlson, C. S., Casey, G., Chan, A. T., Chang-Claude, J., Chanock, S. J., Chen, L. S., Coetzee, G. A., ... Hsu, L. (2013). Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. Gastroenterology, 144(4), 799–807. https://doi.org/10.1053/j.gastro.2012.12.020

Pfaff, C. L., Parra, E. J., Bonilla, C., Hiester, K., McKeigue, P. M., Kamboh, M. I., Hutchinson, R. G., Ferrell, R. E., Boerwinkle, E., & Shriver, M. D. (2001). Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. American Journal of Human Genetics, 68(1), 198–207. https://doi.org/10.1086/316935

Pharoah, P. D. P., Antoniou, A. C., Easton, D. F., & Ponder, B. A. J. (2008). Polygenes, Risk Prediction, and Targeted Prevention of Breast Cancer. New England Journal of Medicine, 358(26), 2796–2803. https://doi.org/10.1056/nejmsa0708739

Pischon, T., Nöthlings, U., & Boeing, H. (2008). Obesity and cancer. Proceedings of the Nutrition Society, 67(2), 128–145. https://doi.org/10.1017/S0029665108006976

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38(8), 904–909. https://doi.org/10.1038/ng1847

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. Nature Reviews Genetics, 12(8), 529–541. https://doi.org/10.1038/nrg3000

Rodríguez Sala-Gómezgil, M. L. (1983). El lenguaje como elemento cultural de identidad social en la zona fronteriza del norte de México. Estudios Fronterizos, 2, 153–178. https://doi.org/10.21670/ref.1983.02.a06

Rojas, W., Parra, M. V., Campo, O., Caro, M. A., Lopera, J. G., Arias, W., Duque, C., Naranjo, A., García, J., Vergara, C., Lopera, J., Hernandez, E., Valencia, A., Caicedo, Y., Cuartas, M., Gutiérrez, J., López, S., Ruiz-Linares, A., & Bedoya, G. (2010). Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. American Journal of Physical Anthropology, 143(1), 13–20. https://doi.org/10.1002/ajpa.21270

Rubio Badán, J. C. (2014). Censos y población indígena en México: algunas reflexiones.Comisión Económica para América Latina y el Caribe (Naciones Unidas (Ed.);CEPAL). https://www.cepal.org/es/publicaciones

Ruiz-Linares, A., Adhikari, K., Acuñ A-Alonzo, V., Quinto-Sanchez, M., & Jaramillo, C. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. PLoS Genet, 10(9), 1004572. https://doi.org/10.1371/journal.pgen.1004572

Salzano, F. M., & Sans, M. (2014). Interethnic admixture and the evolution of Latin American populations. Genetics and Molecular Biology, 37(1), 151–170. https://doi.org/10.1590/S1415-47572014000200003

Sandoval, J. R., Salazar-Granara, A., Acosta, O., Castillo-Herrera, W., Fujita, R., Pena, S. D., & Santos, F. R. (2013). Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. Journal of Human Genetics. https://doi.org/10.1038/jhg.2013.73

Sandoval, K., Moreno-Estrada, A., Mendizabal, I., Underhill, P. A., Lopez-Valenzuela, M., Peñaloza-Espinosa, R., Lopez-Lopez, M., Buentello-Malo, L., Avelino, H., Calafell, F., & Comas, D. (2012). Y-chromosome diversity in Native Mexicans reveals continental transition of genetic structure in the Americas. American Journal of Physical Anthropology, 148(3), 395–405. https://doi.org/10.1002/ajpa.22062

Sans, M. (2000). Admixture studies in Latin America: from the 20th to the 21st century. Human Biology, 72(1), 155–177 https://www.jstor.org/stable/41465813

Sans, M., Salzano, F. M., & Chakraborty, R. (1997). Historical Genetics in Uruguay: Estimates of Biological Origins and Their Problems. Human Biology, 69(2), 161–170.https://www.jstor.org/stable/41435808

Sans, M., Weimer, T., Franco, M. H., Salzano, F. M., Bentancor, N., Alvarez, I., Bianchi, N. O., & Chakraborty, R. (2002). Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. American Journal of Physical Anthropology, 118(1), 33–44. https://doi.org/10.1002/ajpa.10071

Schmit, S. L., Schumacher, F. R., Edlund, C. K., Conti, D. V., Ihenacho, U., Wan, P., Van Den Berg, D., Casey, G., Fortini, B. K., Lenz, H.-J. J., Tusié-Luna, T., Aguilar-Salinas, C. A., Moreno-Macías, H., Huerta-Chagoya, A., Ordóñez-Sánchez, M. L., Rodríguez-Guillén, R., Cruz-Bautista, I., Rodríguez-Torres, M., Muñóz-Hernández, L. L., ... Figueiredo, J. C. (2016). Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis, 37(6), 547–556. https://doi.org/10.1093/carcin/bgw046

Schubert S.A., Morreau H., de Miranda N.F.C.C & van Wezel T. (2020) The missing heritability of familial colorectal cancer. Mutagenesis 35,221–231. https://doi.org/10.1093/mutage/gez027

Service, S., DeYoung, J., Karayiorgou, M., Roos, J. L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J. A., Heutink, P., Aulchenko, Y., Oostra, B., Van Duijn, C., Jarvelin, M. R., Varilo, T., Peddle, L., Rahman, P., Piras, G., ... Freimer, N. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nature Genetics, 38(5), 556–560. https://doi.org/10.1038/ng1770

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. Nucleic Acids Research, 29(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., & Jimenez-Sanchez, G. (2009). Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proceedings of the National Academy of Sciences, 106(21), 8611–8616. https://doi.org/10.1073/pnas.0903045106

Smith, M. W., Patterson, N., Lautenberger, J. a, Truelove, A. L., McDonald, G. J., Waliszewska, A., Kessing, B. D., Malasky, M. J., Scafe, C., Le, E., De Jager, P. L., Mignault, A. a, Yi, Z., De The, G., Essex, M., Sankale, J.-L., Moore, J. H., Poku, K., Phair, J. P., ... Reich, D. (2004). A high-density admixture map for disease gene discovery in african americans. American Journal of Human Genetics, 74(5), 1001–1013. https://doi.org/10.1086/420856

Tenesa, A., Farrington, S., Prendergast, J., Porteous, M. E., Walker, M., Haq, N., Barnetson, R., & Theodoratou, E. (2008). Genome-wide asoociation scan identifies a colorectal cancer suceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Human Genetics, 40(5), 631–637. https://doi.org/10.1038/ng.133.Genome-wide

The Human Genome Project. (2020). https://www.genome.gov/human-genome-project

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A. M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., Jaeger, E., Fielding, S., Rowan, A., Vijayakrishnan, J., Domingo, E., Chandler, I., Kemp, Z., Qureshi, M., Farrington, S. M., ... Houlston, R. S. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nature Genetics, 40(5), 623–630. https://doi.org/10.1038/ng.111

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., ... Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nature Genetics, 39(8), 984–988. https://doi.org/10.1038/ng2085

Verkasalo, P. K., Kaprio, J., Koskenvuo, M., & Pukkala, E. (1999). Genetic predisposition, environment and cancer incidence: A nationwide twin study in Finland, 1976-1995. International Journal of Cancer, 83(6), 743–749. https://doi.org/10.1002/(SICI)1097-0215(19991210)83:6<743::AID-IJC8>3.0.CO;2-Q

Wang, H., Burnett, T., Kono, S., Haiman, C. A., Iwasaki, M., Wilkens, L. R., Loo, L. W. M., Van Den Berg, D., Kolonel, L. N., Henderson, B. E., Keku, T. O., Sandler, R. S., Signorello, L. B., Blot, W. J., Newcomb, P. A., Pande, M., Amos, C. I., West, D. W., Bézieau, S., ... Le Marchand, L. (2014). Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nature Communications, 5. https://doi.org/10.1038/ncomms5613

Wang, H., Haiman, C. A., Burnett, T., Fortini, B. K., Kolonel, L. N., Henderson, B. E., Signorello, L. B., Blot, W. J., Keku, T. O., Berndt, S. I., Newcomb, P. A., Pande, M., Amos, C. I., West, D. W., Casey, G., Sandler, R. S., Haile, R., Stram, D. O., & Le Marchand, L. (2013). Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. Human Molecular Genetics, 22(24), 5048–5055. https://doi.org/10.1093/hmg/ddt337

Wang, H., Schmit, S. L., Haiman, C. A., Keku, T. O., Kato, I., Palmer, J. R., van den Berg, D., Wilkens, L. R., Burnett, T., Conti, D. V., Schumacher, F. R., Signorello, L. B., Blot, W. J., Zanetti, K. A., Harris, C., Pande, M., Berndt, S. I., Newcomb, P. A., West, D. W., ... González-Villalpando, M. E. (2017). Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. International Journal of Cancer, 140(12), 2728–2733. https://doi.org/10.1002/ijc.30687

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V, Molina, J. a, Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., ... Ruiz-Linares, A. (2007). Genetic variation and population structure in native Americans. PLoS Genetics, 3(11), e185. https://doi.org/10.1371/journal.pgen.0030185

Wang, S., Ray, N., Rojas, W., Parra, M. V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A. M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J. A., Freimer, N. B., Bortolini, M. C., Salzano, F. M., Petzl-Erler, M. L., ... Ruiz-Linares, A. (2008). Geographic Patterns of Genome Admixture in Latin American Mestizos. PLoS Genetics, 4(3), e1000037. https://doi.org/10.1371/journal.pgen.1000037

WHO | Noncommunicable diseases: the slow motion disaster. (2017). WHO. http://www.who.int/publications/10-year-review/ncd/en/

World Population Prospects - Population Division - United Nations. (2020). https://population.un.org/wpp/DataQuery/

Xue, Y., Li, C., Duan, D., Wang, M., Han, X., Wang, K., Qiao, R., Li, X. J., & Li, X. L. (2020). Genome-wide association studies for growth-related traits in a crossbreed pig population. Animal Genetics. https://doi.org/10.1111/age.13032

Ye, P., Xi, Y., Huang, Z., & Xu, P. (2020). Linking obesity with colorectal cancer: Epidemiology and mechanistic insights. In Cancers (Vol. 12, Issue 6). MDPI AG. https://doi.org/10.3390/cancers12061408

Zanke, B. W., Greenwood, C. M. T., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E., Ferretti, V., Laflamme, P., Sundararajan, S., Roumy, S., Olivier, J. F., Robidoux, F., Sladek, R., Montpetit, A., Campbell, P., ... Dunlop, M. G. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nature Genetics, 39(8), 989–994. https://doi.org/10.1038/ng2089

Zeng, C., Matsuda, K., Jia, W. H., Chang, J., Kweon, S. S., Xiang, Y. B., Shin, A., Jee, S. H., Kim, D. H., Zhang, B., Cai, Q., Guo, X., Long, J., Wang, N., Courtney, R., Pan, Z. Z., Wu, C., Takahashi, A., Shin, M. H., ... Thomas, D. C. (2016). Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. Gastroenterology, 150(7), 1633–1645. https://doi.org/10.1053/j.gastro.2016.02.076

Zhang, B., Jia, W. H., Matsuda, K., Kweon, S. S., Matsuo, K., Xiang, Y. B., Shin, A., Jee, S. H., Kim, D. H., Cai, Q., Long, J., Shi, J., Wen, W., Yang, G., Zhang, Y., Li, C., Li, B., Guo, Y., Ren, Z., ... Zheng, W. (2014). Large-scale genetic study in east Asians identifies six new loci associated with colorectal cancer risk. Nature Genetics, 46(6), 533–542. https://doi.org/10.1038/ng.2985

Zorcolo, L., Fantola, G., Balestrino, L., Restivo, A., Vivanet, C., Spina, F., Cabras, F., Ambu, R., & Casula, G. (2011). MUTYH-associated colon disease: adenomatous polyposis is only one of the possible phenotypes. A family report and literature review. Tumori, 97(5). https://doi.org/10.1700/989.10731

Capítulo 7

Anexo I

7.1. Artículo publicado en *Genetics and Molecular Biology*

Genetics and Molecular Biology 43, 4, e20200143 (2020)

DOI: https://doi.org/10.1590/1678-4685-GMB-2020-0143

Research Article
Human and Medical Genetics

Differential admixture in Latin American populations and its impact on the study of colorectal cancer

Valentina Colistro[1], Patricia Mut[2], Pedro C. Hidalgo[3], Angel Carracedo[4,5], Inés Quintela[4], Augusto Rojas-Martínez[6] and Mónica Sans[2]

[1]*Universidad de la República, Facultad de Medicina, Departamento de Métodos Cuantitativos, Montevideo, Uruguay.*
[2]*Universidad de la República, Facultad de Humanidades y Ciencias de la Educación, Departamento de Antropología Biológica, Montevideo, Uruguay.*
[3]*Universidad de la República, Centro Universitario de Tacuarembó, Polo de Desarrollo Universitario Diversidad Genética Humana, Tacuarembó, Uruguay.*
[4]*Universidad de Santiago de Compostela, Centro Nacional de Genotipado (CEGEN), Spain.*
[5]*Universidade de Santiago de Compostela, CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud Carlos III, Grupo de Medicina Xenómica, Santiago de Compostela, Spain.*
[6]*Escuela de Medicina y Ciencias de la Salud, Tecnológico de Monterrey, Monterrey, México.*

Abstract

Genome-wide association studies focused on searching genes responsible for several diseases. Admixture mapping studies proposed a more efficient alternative capable of detecting polymorphisms contributing with a small effect on the disease risk. This method focuses on the higher values of linkage disequilibrium in admixed populations. To test this, we analyzed 10 genomic regions previously defined as related with colorectal cancer among nine populations and studied the variation pattern of haplotypic structures and heterozygosity values on seven categories of SNPs. Both analyses showed differences among chromosomal regions and studied populations. Admixed Latin-American samples generally show intermediate values. Heterozygosity of the SNPs grouped in categories varies more in each gene than in each population. African related populations have more blocks per chromosomal region, coherently with their antiquity. In sum, some similarities were found among Latin American populations, but each chromosomal region showed a particular behavior, despite the fact that the study refers to genes and regions related with one particular complex disease. This study strongly suggests the necessity of developing statistical methods to deal with di- or tri-hybrid populations, as well as to carefully analyze the different historic and demographic scenarios, and the different characteristics of particular chromosomal regions and evolutionary forces.

Keywords: Admixture, genetic ancestry, heterozygosity, Latin American populations.

Received: May 07, 2020; Accepted: September 14, 2020.

Introduction

One of the greatest challenges in genetic epidemiology is the development and application of methodological strategies allowing identification of genetic risk loci in order to achieve a more thorough understanding of the genetic basis of complex diseases, as they are the result of interactions between multiple genetic and/or environmental factors, each with modest effects. It is likely that different combinations produce the same clinical symptoms (Botstein and Risch, 2003). Also, many complex diseases are genetically related, sharing common genetic risk variants (Teng *et al.*, 2016). Moreover, interconnections among all genes expressed in disease-relevant cells and the core disease-related genes ("omnigenic" model) (Boyle *et al.*, 2017).

Linkage and association studies are the two main approaches applied to identify the genetic basis of these types of diseases (Morton, 2003; Patel *et al.*, 2003; Khoury *et al.*, 2010). Linkage studies are more efficient in detecting genes with large effects, like single-gene based disorders, but they lack the statistical power to detect variants with modest effects. On the other hand, genome-wide association studies (GWAS) have a statistical advantage as they provide greater power for detecting common variants with modest risk (Risch and Merikangas, 1996). However, these studies have been criticized, as they rely on an extremely high number of markers in order to be carried out (more than 100.000), a large quantity of samples, as well as adequate technological resources to process the enormous amount of data, becoming impractical and very expensive (Cantor *et al.*, 2010; Qin and Zhu, 2012).

Send correspondence to Mónica Sans. Departamento de Antropología Biológica, Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Magallanes 1577, 11200 Montevideo, Uruguay. E-mail: mbsans@gmail.com.

Admixture mapping studies (AMSs) constitute an alternative approach. This methodology was first proposed by Rife (1953), but its implementation has been technically possible only in the last decades (McKeigue, 2005). AMS is based on the gene flow processes between continental populations occurring in the last centuries, producing particular chromosome configurations in the resulting admixed populations, showing a mosaic of ancestry segments (Darvasi and Shifman, 2005). When a disease has substantial despair prevalence among parental populations, the risk allele locus will show an over-representation ancestry of the high risk population in the admixed population. The use of ancestry informative markers (AIMs) allows the identification of the population source of the studied chromosomal segments (Tian *et al.*, 2008; Winkler *et al.*, 2010, among others). The effect of rare variants in recently admixed populations can be much greater compared with its ancestral populations, as has been shown by Moltke and Albrechtsen (2014). Moreover, the effects of noncausal genetic variants depend on its correlation with causal variants, and these last may vary depending on the ancestral populations and the patterns of linkage disequilibrium (Skotte *et al.*, 2019).

The process of admixture in the Americas can be seen as a natural experiment for genetic epidemiology and anthropology, in which polymorphic marker loci are used to infer a genetic basis for traits of interest (Chakraborty and Weiss, 1988). Nowadays it is possible to establish a maximum of approximately 21 generations of admixing, depending on the region. Cosmopolitan Latin American populations have Native contributions from around 1% to more than 50%, and African contributions from 2 to 40%, while on the other side, it is rare to find Native groups without any admixture (Sans, 2000).

The grade of contribution of each parental population will be reflected not only in the amount of chromosomes from each ancestral origin, but in the quantity of blocks from these origins inside chromosomes, and depend on the admixture process (Pfaff *et al.*, 2001). We assume that the antiquity of populations is directly related to the heterozygosity and the size of chromosomal blocks; consequently, we expect smaller blocks in more ancient populations. Moreover, heterozygosity can be related to the time (generations) after a process of admixture, assuming that non-admixed populations are more homogeneous. We recognize that it is an oversimplification because it ignores the microevolutionary changes in the admixed population, as genetic drift, selection and gene flow.

The major aim of our study was to understand the process that generates complex chromosome patterns in admixed populations and to improve the implementation of AMSs in Latin American populations. Particularly, this study was focused on analyzing genes and chromosomal regions previously related to colorectal cancer (CRC) in admixed populations, because past studies were mainly based on populations of European descent. CRC is common in both sexes and has no major avoidable risk factor. By determining the ancestral proportions, as well as the heterozygosity and size of fragments in five admixed American populations and several populations from Europe, Africa and Asia in associated regions, we intend to help in the understanding of genetic CRC causes.

Subjects and Methods

Samples

We used data available in 1000 Genomes Project for eight populations and an unpublished set of genetically admixed Mexican samples. Regarding the 1000 Genome Project samples (The 1000 Genomes Consortium, 2010), five are admixed populations from the Americas, and the others were selected to represent part of their parental populations. The admixed populations were: Afro-Americans from the United States (ASW, N=83), Colombians (CLM, N=60), Puerto Ricans (PUR, N=55), Peruvians (PEL, N=85) and Mexicans from Los Angeles, CA (MXL, N=76). The samples from Africa, Europe and Asia were selected due to their relationship to the migrations toward America, being the last ones considered in substitution of Native Americans. We are aware of differences between Asian and Native American populations, but we choose this alternative due to the scarcity of data referred to the SNPs and regions considered for such populations. Therefore, we analyzed Yorubas and Luhya to represent African populations (denominated Africans, AF, in this study, N=176), Iberians, Tuscans, and Utah residents with northern and western European ancestry for European populations (denominated EU, N=174), and Chinese from Beijing, Southern Han Chinese and Japanese from Tokyo to represent Asians (denominated AS, N=98).

We are particularly interested in another Mexican sample (hereafter, MEX, N=831) because it is formed by healthy controls of a GWAS study of CRC (CHIBCHA, study of hereditary cancer in Europe and Latin America). The individuals were recruited in different blood banks, three in Mexico City (Centro Médico Nacional Siglo XXI of the Mexican Social Security Institute -IMSS), three in Monterrey (UMAE 25, IMSS and the University Hospital of the Universidad Autónoma de Nuevo León) and three in Torreon (UFM 16 IMSS, the UMAE 71 IMSS, and the University Hospital of Torreon), from 2010 to 2012. All subjects gave informed consent for inclusion before they participated in the study. The protocol was approved by the ethics committees of each participating institution (Ethics Committee of the University Hospital, Universidad Autónoma de Nuevo León code BI10-003 and the National Commission of Scientific Research of the Mexican Social Security Institute code R-2012-785-032), the Federal Commission for Protection against Health Risks (COFEPRIS), code CMN2012-001, and the Ethics Committee of CHIBCHA project number: 223 678.

Samples were genotyped using two complementary arrays: Axiom Genome-Wide LAT 1 (Latino) Array and a Custom-designed Array, both from Affymetrix Axiom Genotyping Solutions. The former was designed to maximize coverage of common and rare disease-associated alleles in Latin American populations that have genetic contri-

butions from European, Native American and African ancestries. The latter was specifically designed for this study, being the SNPs selection based on regions previously detected as associated with CRC. SNP calling in both arrays was done following Affymetrix best practice workflow, which includes the Genotyping Console Software in combination with SNPolisher. A total of 1,169,944 SNPs (387,948 from the Custom Array and 781,996 from the Latino Array) was obtained. These samples were included because its large number of individuals and the high coverage of SNPs in the considered regions represent an opportunity to compare the performance of another admixed population.

Genotypes of Native American (NAM) samples were used in order to estimate the global individual ancestry. These genotypes included individuals from five ethnic groups: Zapotecs from Oaxaca, Mexico (N=21), Tepehuans from Durango in Northern Mexico (N=23), Nahuas from Central Mexico (N=14), Mayas from Campeche, Mexico (N=25), Quechuas from Cerro de Pasco, Perú (N=24) and Aymaras from La Paz, Bolivia (N=25). We consider a panel of AIMs developed and optimized for the study of Latin American populations by the LACE Consortium (for detailed information about the panel and the populations refer to Galanter *et al.*, 2012). This panel was composed of 446 AIMs but the ancestry analysis performed in the present study was limited to the 275 SNPs shared with the Mexicans, the 1000G populations and the Native American samples.

Genomic regions studied

We selected 10 autosomal regions, with an average size of 680.9 Kbp spanning a total of 6.8 MB (Table 1). These regions were previously described to show association to CRC (Kinzler *et al.*, 1991; Aaltonen *et al.*, 2007), seven of them are genes: *APC*, *BRAF*, *MSH2*, *MSH6*, *MLH1*, *MUTYH* and *PMS2*, and three are loci described by Carvajal-Carmona *et al.* (2011) also associated with CRC: 8q23.3, 16q22.1 and 19q13.11.

For the seven gene regions, SNPs within the gene limits were retrieved, and in the three other regions, 1 MB upstream and downstream SNPs were considered. The number of available SNPs in each region is listed in Table 1.

Admixture analysis

In order to understand the structure of the MEX sample, we performed a global individual admixture analysis using the AIMs panel described above. Estimation of individual admixture fractions were calculated with ADMIXTURE software version 1.3.1 (Alexander *et al.*, 2009), which considers a likelihood model. To choose the correct value of k we computed the cross-validation error over k, from 2 to 6. We found that k=3 yielded the lowest cross-validation error (k_3=0.538) compared to other k values (k_2=0.63968, k_4=0.54016, k_5=0.54226 and k_6=0.542).

Complementary, we also analyzed the mean population admixture in each of the 10 regions for the admixed populations. In this case we were not able to use the Native American samples due to their limited number of SNPs yielding at these 10 regions. As explained above, we used the Asian samples instead. A total of 5283 SNPs were used for this analysis.

Analysis of genetic variation

The genetic variation analysis was performed only on the seven genomic regions corresponding to genes. To compare the variation in the studied regions among the nine populations, we considered two measures using PLINK version 1.9 (Purcell *et al.*, 2007; Chang *et al.*, 2015): heterozygosity and haplotypic structures among regions and populations.

For the heterozygosity determination, the mean values of heterozygosity were analyzed for each gene by population and the mean values of SNPs were classified in seven categories. The SNPs classification categories are related to their position and consequence to transcript and were obtained using Biomart (Haider *et al.*, 2009): intronic, non-synonymous coding, synonymous coding, 5' UTR, 3' UTR, stop gained and stop lost.

Inference of haplotype phase was determined with the Beagle software version 4 (Browning and Browning, 2007). Gabriel *et al.* (2002) criteria were followed to define haploblocks. The allelic association between pairs of SNPs was measured by the D' parameter (Lewontin, 1964). The distribution of blocks length (in bp) among populations was compared. Linkage analysis and haploblock estimation were done using PLINK version 1.9 (Purcell *et al.*, 2007).

Table 1 - The 10 genomic regions considered in the analysis, 7 were genes and 3 were locations (± 1 MB). The table shows chromosome, base pair start and end, gene name, cytoband and number of SNPs of each studied location.

Chromosome	Band	Gene Start (bp)	Gene End (bp)	Gene Name	SNPs
1	p34.1	45794835	45806142	MUTYH	706
2	p21	47630108	47789450	MSH2	1058
2	p16.3	48010221	48034092	MSH6	1011
3	p22.2	37034823	37107380	MLH1	997
5	q22.2	112043195	112181936	APC	1140
7	q34	140424943	140624564	BRAF	1138
7	p22.1	6012870	6048756	PMS2	682
8	q23.3	116631278	118626279	——	864
16	q22.1	67824395	69816284	——	964
19	q13.11	32534093	34530086	——	1025
Total					9585

Results

Admixture analysis

The AIM panel accurately discriminates parental populations, as can be seen in Figure 1a. The representation of the global individual ancestral fractions for the admixed populations is shown in Figure 1b and 1c. According to the estimations, the ASW population has 75,4% of African ancestry, while the African proportions for the other admixed populations were lower: 12,1% in CLM, 6,8% in MXL, 4,3% in PEL and 16% in PUR. Peruvian samples (PEL) have the highest proportions of Native American ancestry (77,1%) followed by the Mexican samples (MXL and MEX) (51,2 and 61,5%, respectively). The European ancestry has its maximum in Puerto Rico (68,7%) followed by the Colombian sample (61,7%) (Table 2).

When the admixture analysis was performed on those 10 regions considered in this study, the results show high variability (Figure 2). No clear pattern is detected among the different regions. In general terms, there is a greater concordance among populations than among genes and regions. The greatest similarity is between both Mexican samples, while Peruvians seems to be the most dissimilar. While in MSH6 and MLH1 genes, a greater contribution of Asian ancestry was detected, and in 16q22.1 and MUTYH the European contribution is the highest.

Genetic variation

The results of the analyses of the mean heterozygosity by gene are shown in Table 2 and the mean heterozygosity using the categories of SNPs mentioned above are shown in Figure 3. For two of these categories (stop gained, stop lost), no population showed heterozygosity in any region.

The greatest mean values of heterozygosity for most of the genes are found in the ASW, except for BRAF, MSH6 and MUTYH where the greatest values are in AF, EU and AS respectively. And the lowest values are found in AS for MLH1, MSH2 and MSH6; in PEL for APC and MUTYH, in MEX for BRAF and in MXL for PMS2 (Table 2).

When including the SNP category in the analysis, different genes show different situations: a) heterozygosity related to categories of SNPs vary in different regions; b) some chromosomal regions do not show heterozygosity in some categories of SNPs; c) heterozygosity varies when considering different populations, but its behavior is relatively coherent in the different categories: Africans and Afro-descen-

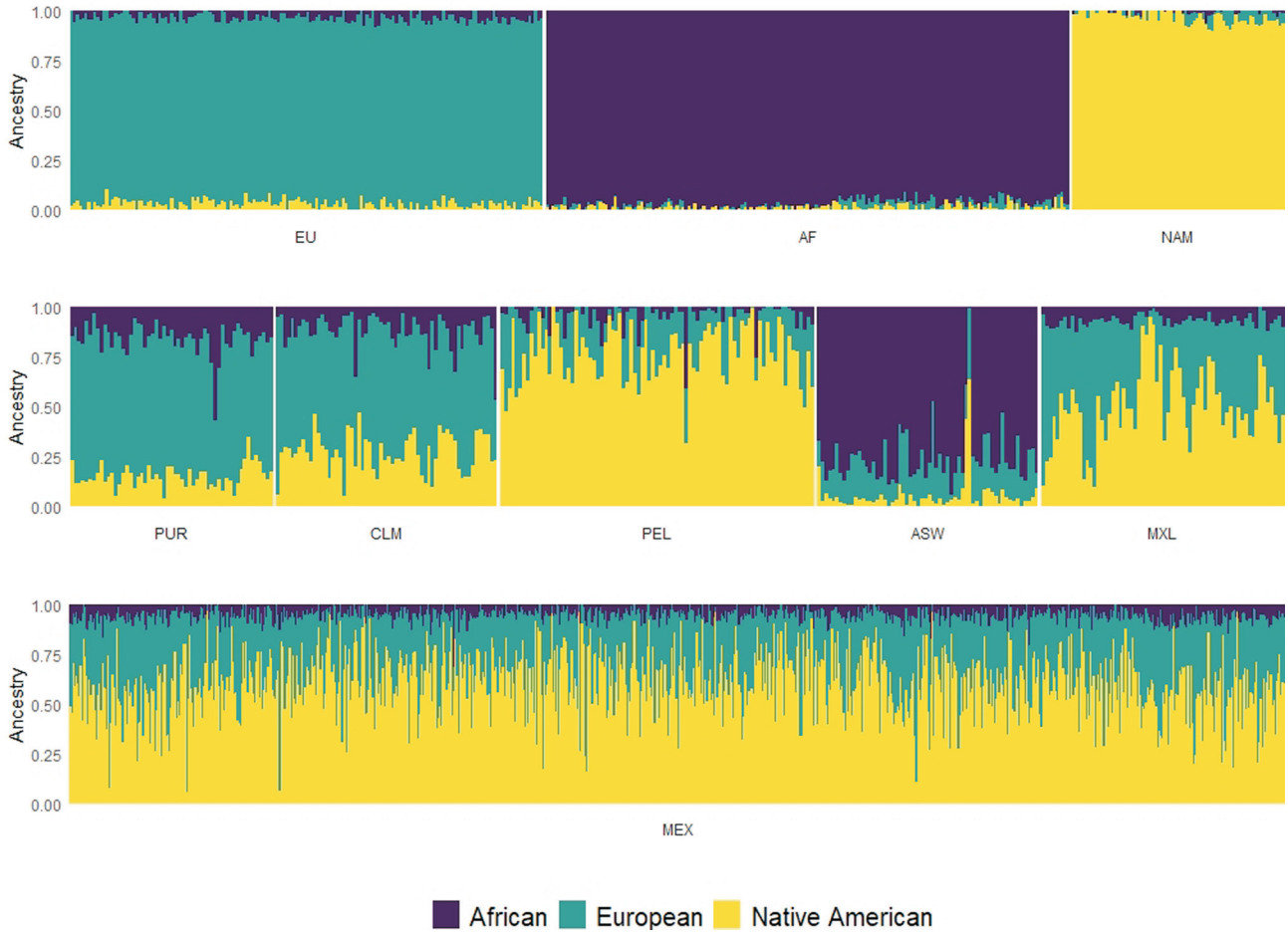

Figure 1 - Global admixture analysis performed in ADMIXTURE, with k=3 representing the 3 ancestral components of the Admixed American populations. The barplots show each individual as a vertical line, and the ancestries are indicated by different color (NAM= Native American ancestry, AFR= African ancestry and EUR= European ancestry). a) Parental populations, b) Admixed populations from 1000G and c) Mexican unpublished samples.

Table 2 - Mean values of heterozygosity by population and by region. Highest and lowest values of each row were marked in order to facilitate visualization.

Gene	AF	ASW	AS	EU	PUR	CLM	PEL	MEX	MXL
APC	0,063	0,066*	0,046	0,056	0,049	0,047	0,0378†	0,044	0,044
BRAF	0,079*	0,080	0,046	0,036	0,048	0,041	0,031	0,027†	0,029
MLH1	0,080	0,094*	0,016†	0,059	0,052	0,052	0,052	0,057	0,061
MSH2	0,063	0,065*	0,049†	0,052	0,054	0,061	0,052	0,052	0,054
MSH6	0,053	0,051	0,025†	0,082*	0,079	0,078	0,035	0,047	0,049
MUTYH	0,031	0,026	0,039*	0,024	0,032	0,032	0,020†	0,026	0,028
PMS2	0,094	0,101*	0,079	0,080	0,080	0,071	0,077	0,075	0,068†
Total	0,071	0,073*	0,045	0,052	0,054	0,051	0,042†	0,044	0,045

* highest value in row; † lowest value in row

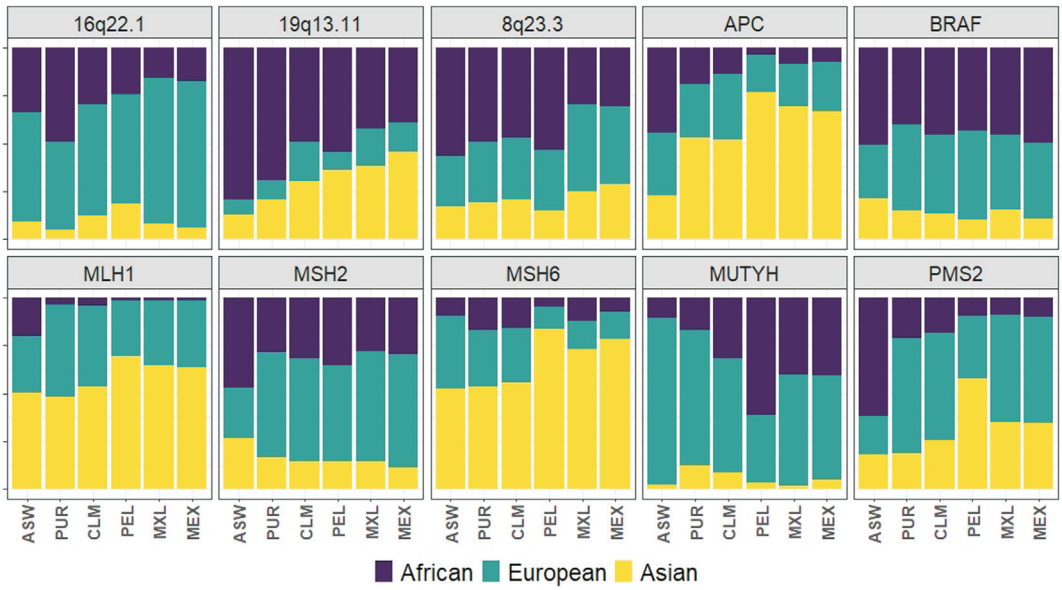

Figure 2 - Admixture analysis by region performed with ADMIXTURE, with k=3 representing the 3 ancestral components of the Admixed American populations. The barplots show the mean ancestry of each population, and the ancestry proportions are indicated by different colors.

dants, European and Asian, and Latin American admixed ones.

For example, for the *APC* gene, relative differences in heterozygosity values associated to the categories of SNPs remain constant in all populations being 3'UTR the one with greatest values of heterozygosity followed by synonymous variants, except in the EU and CLM samples, where synonymous variants are greater (Figure 3).

Regarding the *BRAF* gene, the diversity among populations is clear. For this gene, the African related samples (AFR and ASW) have higher values of heterozygosity in intronic and synonymous categories, while 3'UTR regions are more homogeneous. Puerto Rico has an intermediate place between African related and other considered populations (Figure 3).

The *MSH2* locus differs from the others analyzed. The 3'UTR SNPs show none or very small heterozygosity in every population, except for the AF, ASW and PUR samples. As populations and MXL do not have heterozygosity in 3'UTR and synonymous mutations, while MEX shows very little heterozygosity in those regions. Intronic SNPs show the higher heterozygosity in every population (Figure 3).

It is important to note that the admixed Latin-American samples (PEL, CLM, MXL, MEX and PUR) show heterozygosity values for all genes that tend to be intermediate among the values of the parental samples (EU, AS and AF). Although, the ASW, also admixed, shows a pattern closer to the AF than to any other sample, in concordances with the high contributions of African genes (76%); in some cases, also PUR approximates more to those samples.

The quantity of phased haplotype blocks per gene was analyzed for each population (Figure 4). The African populations (AF and ASW) have more blocks per region for most of the genes, while Asians (AS) have fewer, followed by MXL, probably because of the high Native American contribution of Native genes (62%), and by Europeans. All populations have a similar curve for the 7 genes, with some exceptions: CLM shows a large amount of haplotype blocks in *BRAF*, PUR that shows more blocks in *MSH6*, and MEX that shows more blocks at *MSH2* and *PMS2* genes and an unexpected behavior related to the other Mexican sample (MXL).

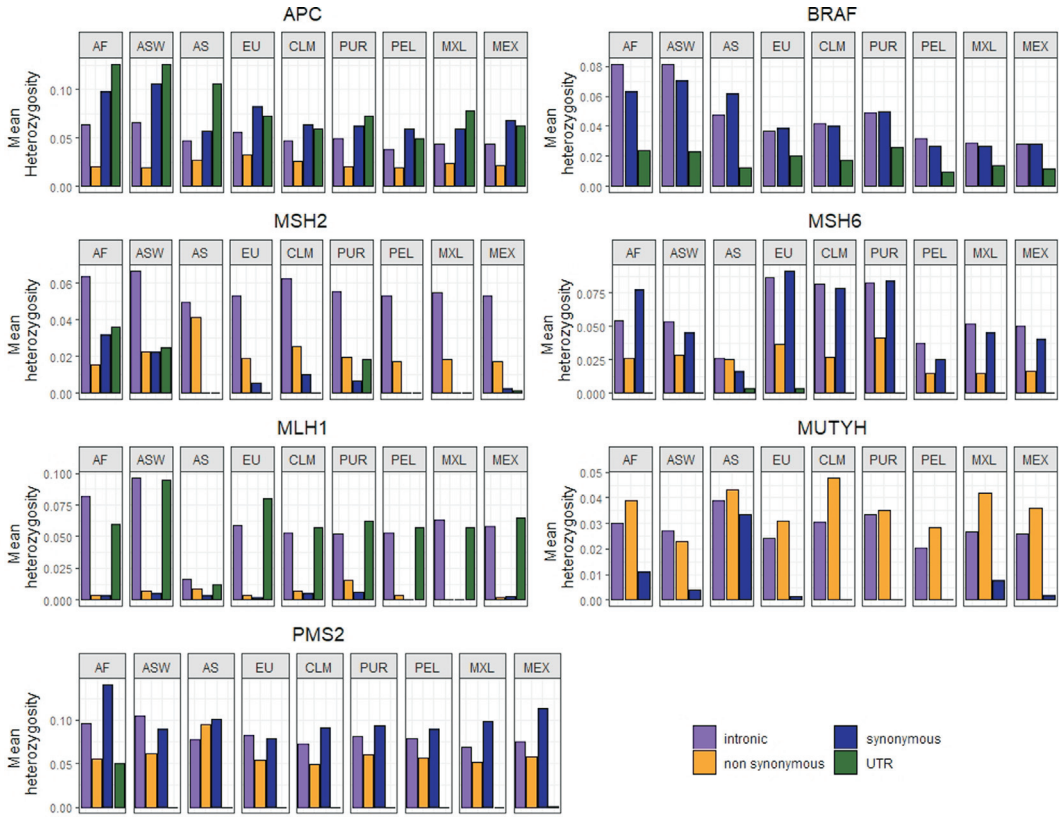

Figure 3 - Mean heterozygosity values in four SNPs categories by gene and population. Each bar corresponds to a SNP category in a certain gene and population. SNPs categories are: intronic, non-synonymous, synonymous and UTR.

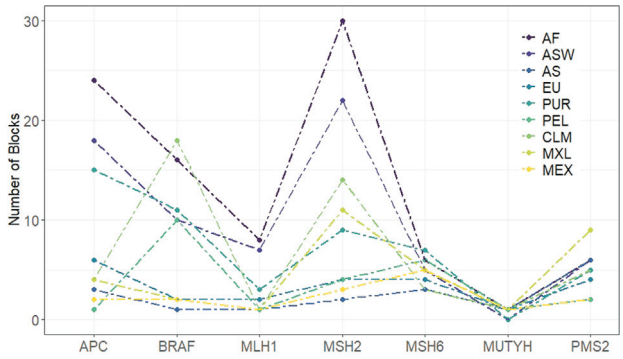

Figure 4 - The graphic shows the number of phased haplotype blocks estimated for the 7 genes detailed per population.

The variability of the size of the blocks shows diversity among populations (Figure 5). It varied from < 1 kb to over 190 kb, though most of the blocks were small (< 5 kb). Markedly the African related populations (AF and ASW) have higher proportions of small blocks, and the admixed populations (CLM, MXL, MEX and PUR) are placed in an intermediate value between the African-related and the other two parental populations (AS and EU). In the MEX sample, the smaller blocks are underrepresented in comparison with the rest of the admixed samples, while they show a greater number of longer blocks related to the rest of admixed populations.

Discussion

The results obtained using the selected AIMs supports the use of these markers for detecting admixture in Latin American populations, as demonstrated in several studies performed before (Mao *et al.*, 2007; Halder *et al.*, 2008; Tian *et al.*, 2008; Silva *et al.*, 2010; Galanter *et al.*, 2012; Manta *et al.*, 2013). Moreover, we found that the expected proportions of ancestry are consistent with the historical and geographical affinities of the samples used, as well as other estimations (Norris *et al.*, 2018). Peruvian and both Mexican samples showed the highest Native contribution, being 77,1% in PUR, 51,2% in MXL and 61,5% in MEX; all three samples have the lowest African one (4-6%). Different studies about population admixture in Mexico showed different contributions. In the central and northern regions, Native American contribution goes from 32% to 69% Native-American while African is usually less than 7% (Martinez-Fierro *et al.*, 2009; Salzano and Sans, 2014). A comprehensive analysis by Rubi-Castellanos *et al.* (2009) in 10 Mexican regions shows somehow different results, presenting higher African contributions in some regions as Nueva Leon (18,5%), Veracruz (17,2%), and Jalisco and Campeche (15,9%).

The ancestry analysis by region evidenced a different result in each one of the 10 regions. While in some genes the Asian contribution (as a proxy of Native American) predominates in all the admixed samples (*MLH1* and *MSH6*), in the 16q22.1 region the European contribution prevails. However, in most of the regions, the predominant ancestry is not

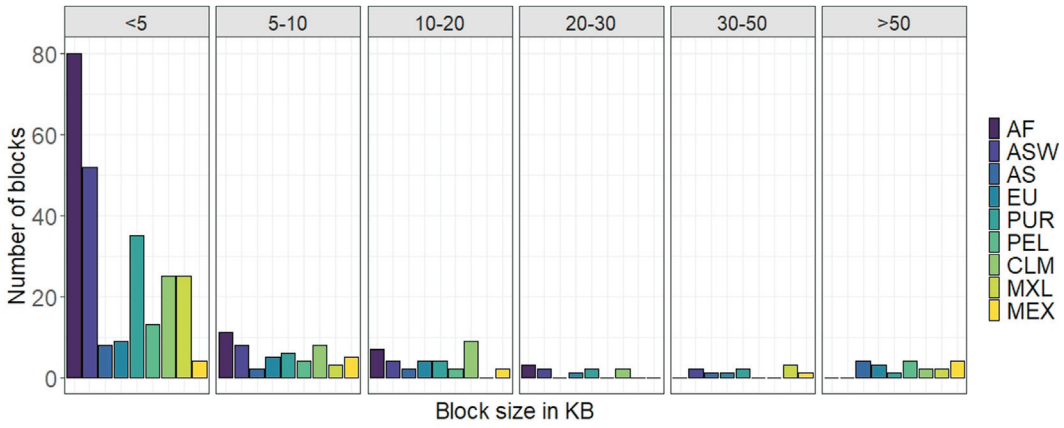

Figure 5 - Block characteristics size (in kb) distribution of all haplotype blocks found in the analysis. Summary of haplotype diversity across all blocks.

the same for all samples. In *MSH2*, the European contribution is predominant except in the ASW in which the African is the greatest. This exposes a different situation for each population and for each genomic region and outlines the importance of considering the local ancestry complementary to the global ancestry when performing association analysis in order to avoid spurious associations.

A similar conclusion can be drawn by taking into account the genetic variation analyses. The heterozygosity values showed very dissimilar ancestral contribution by population and by region. Only in one of the regions considered the highest and lowest mean values of heterozygosity were detected in one parental population (*MSH6*), being in most of the cases the highest mean value found in the African samples (all except *MUTHY* and *MSH6*). And finally, in four cases (*APC, BRAF, MUTHY* and *PMS2*), the lowest mean values were found in two admixed populations (PEL and MXL).

Both, in ancestry and in genetic variation analyses, the Native American contribution in Peruvian and Mexican samples is the highest, and consequently, it is possible to presuppose that the genetic variation patterns could be more closely related to Native Americans than in other Latin American populations. This is reflected in the *MSH2* gene heterozygosity values, as well as for haplotypic blocks of 5-10 kb, but not for the rest of the performed analyses. The non-expected values can be explained by different factors, like comparisons with Asian samples, instead of Native American samples. Moreover, some differences between the two Mexican samples were shown. The Mexican (MXL) sample was recruited in Los Angeles, California, and consequently, it can better be compared with Mexican Americans.

The MEX corresponds to the capital city, composed of subjects from the centre of the country, and Monterrey and Torreon, represented by subjects of northern parts of the country. There is also a difference of 10% of Native contribution, being greater in MEX than in MXL. Another crucial difference is the size of the samples (76 *versus* 831, respectively). This fact is not minor, as bigger samples may uncover heterogeneities due to subestructuration. Then, variation in different parameters can be explained because of that, as the apparent presence of variation in heterozygosity at 3' and synonymous not found in MXL but in MEX in *MSH2* gene (Figure 2), or having more longer blocks shown in MEX sample (Figure 4). Also, differences between Mexican samples can be related to the coverage of the DNA analysis, being low for in MXL and high for MEX. It has been demonstrated by Ros-Freixedes *et al.* (2018) that low coverage can generate bias towards the detection of SNPs, showing that concordance with 10X coverage was 90,5% for genotypes and 95,2% for alleles, while with high coverage those values increased to 99,7 and 99,9%, respectively.

The size of blocks supports that admixed populations have higher values of linkage disequilibrium that lead to a specific pattern of haplotypic structures. For example, PUR showed the higher values of European ancestry but despite that, its heterozygosity values are close to EUR for *BRAF* and *MSH2,* but not for *APC* or for haplotypes, where PUR are more similar to the other admixed samples.

Besides African and African-derived populations showed smaller blocks than the other populations, it is necessary to note that all populations analyzed here show a broad range of small blocks indicating little recombination in the regions, most genes, studied. As Gabriel *et al.* (2002) have demonstrated, African and African-American populations have around half of the genome concentrated in blocks of 22 kb or larger. Here we showed an intermediate situation in the Latin American population, despite some differences depending on the degree of admixture (and the origin of the genetic contributions) and the chromosomal region analyzed.

Two facts can be highlighted: 1) several evolutionary forces- not only genetic flow- act on genetic variability; and 2) each region analyzed has special behavior when genetic variation is analyzed, despite all genes and chromosomal regions analyzed.

Related to the first, our data suggest that the patterns of ancestry and variability appear in certain genomic regions and under certain circumstances, but not in others. Different microevolutive forces such as selection, genetic drift, and eventually recombination, conversion and hitchhiking are probably present (Maynard Smith and Haigh, 1974). Moreover, the evolutionary processes act on genetic regions and genes, being selection (positive or negative) the most important, followed by others as mutations (Salzano, 2005). Besides, genetic flow is related to different migrations in the

history of the involved populations that generated differences in populations and subpopulations (Stumpf and Goldstein, 2003; Choudhry *et al.*, 2006). Consequently, a deeper study taking into account historical and demographic scenarios as well as genetic variability is required before trying to make inferences.

Related to the second, the 10 analyzed regions were detected as associated with CRC in European populations (Kinzler *et al.*, 1991; Aaltonen *et al.*, 2007; Carvajal-Carmona *et al.*, 2011). Interestingly, when these regions were considered in the MEX sample when analyzing CRC in controls and patients, none of these genes showed association with the disease; only the 16q22.1 region was detected as associated (unpublished data). We would like to emphasize that our results suggest that not only global ancestry analysis is important when studying the association of genomic regions to a complex disease in admixed populations, but also regional ancestry analysis is advisable to be performed in order to detect an imbalance of ancestral contribution between cases and controls. Otherwise, associations might be the result of the mentioned imbalance rather than the possible implication of that region in the disease considered.

Several authors (among others, Tishkoff and Verrelli, 2003a,b; González Burchard *et al.*, 2005; Coop *et al.*, 2009) have pointed out the importance of evolutionary factors (such as admixture) to understand the genomic structure of populations. Our data support that each population history and each genomic region needs to be studied independently. Consequently, we emphasize the importance of a prospective analysis of ancestral characteristics of the populations to be studied, especially when dealing with the admixed Latin American populations where the di or tri-parental admix model is the most suitable.

Finally, this study strongly suggests the necessity of developing statistical methods to deal with di or tri-hybrid populations. It is also necessary to carefully analyze the different historical and demographic scenarios of each particular population to avoid generalizations, since, considering Latin America as a whole, is more theoretical than real.

Acknowledgments

To the members of the CHIBCHA Consortium Ian Tomlinson (University of Birmingham, UK), Luis Carvajal-Carmona (University of California, Davis, USA), Chris Holmes (University of Oxford, UK), Sergi Castellvi-Bel (Hospital Clinic, Spain), Manuel Teixeira (Portuguese Institute of Oncology, Portugal Magdalena Echeverry (Universidad del Tolima, Colombia) and Rocío Ortíz López (Tecnológico de Monterrey, México), and to the collaborators who take the Mexican samples. To the technician in Santiago de Compostela who collaborated in the genomic analyses. We are especially grateful to the people of Mexico who participated in the study. This research was financed by the Seventh Framework Programme (FP7) of the European Commission, project number 223 678, "Genetic study of Common Hereditary Bowel Cancers in Hispania and the Americas", to Ian Tomlinson.

Conflict of Interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions

VC wrote the original draft of the manuscript, analyzed the data, was responsible for the data curation and conceived and designed the formal análisis; PM analyzed the data, conceived and designed the formal analysis and wrote the draft of the manuscript; PCH conceived and designed the formal analysis and wrote the draft of the manuscript; AC and ARM were responsible for funding acquisition, administrated the project and were in charge of the supervisión; IQ was responsible for the data curation and reviewed the final version of the draft; MS was responsible for funding acquisition, administrated the project was in charge of the supervision, conceived and designed the formal analysis and wrote the original draft of the manuscript. All authors read and approved the final version.

References

Aaltonen LA, Johns L, Järvinen H, Mecklin JP and Houlston R (2007) Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res 13:356-361.

Alexander DH, Novembre J and Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664.

Botstein D and Risch N (2003) Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33:228-237.

Boyle EA, Li YI and Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. Cell 169:1177-1186.

Browning SR and Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Cantor RM, Lange K and Sinsheimer JS (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet 86:6-22.

Carvajal-Carmona LG, Cazier J-B, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J *et al.* (2011) Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: Refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 20:2879-88.

Chakraborty R and Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A 85:9119-9123.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM and Lee JJ (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4:7.

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai H-J, Naqvi M, Phong A, Ung N *et al.* (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118:652-664.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Myers RM, Cavalli-sforza LL, Feldman MW and Pritchard JK (2009) The role of geography in human adaptation. PLoS Genet 5:e1000500.

Darvasi A and Shifman S (2005) The beauty of admixture. Nat Genet 37:118-119.

Gabriel SB, Schaffner SFSF, Nguyen H, Moore JMJM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al.* (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Galanter JM, Fernández-López JC, Gignoux CR, Barnholtz-Sloan J, Fernández-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras AV, Figueroa LU, Raska P *et al.* (2012) Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet 8:e1002554.

González Burchard E, Borrell LN, Choudhry S, Naqvi M, Tsai H-J, Rodriguez-Santana JR, Chapela R, Rogers SD, Mei R, Rodríguez-Cintron W *et al.* (2005) Latino populations: A unique opportunity for the study of race, genetics, and social environment in epidemiological research. Am J Public Health 95:2161-2168.

Haider S, Ballester B, Smedley D, Zhang J, Rice P and Kasprzyk A (2009) BioMart Central Portal–unified access to biological data. Nucleic Acids Res 37:W23-W27.

Halder I, Shriver M, Thomas M, Fernandez JR and Frudakis T (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: Utility and applications. Hum Mutat 29:648-658.

Khoury MJ, Bedrosian SR, Gwinn M, Higgins JPT, Ioannidis JPA and Little J (2010) Human genome epidemiology. 2nd edition. Oxford University Press, New York.

Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D *et al.* (1991) Identification of FAP locus genes from chromosome 5q21. Science 253:661-665.

Lewontin RCC (1964) The interaction of selection and linkage II. Optimum models. Genetics 50:757-782.

Manta FSN, Pereira R, Caiafa A, Silva DA, Gusmão L and Carvalho EF (2013) Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. Ann Hum Biol 40:94-98.

Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, Leon-Velarde F, Moore LG, Vargas E, McKeigue PM *et al.* (2007) A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet 80:1171-1178.

Martinez-Fierro ML, Beuten J, Leach RJ, Parra EJ, Cruz-Lopez M, Rangel-Villalobos H, Riego-Ruiz LR, Ortiz-López R, Martínez-Rodríguez HG and Rojas-Martínez A (2009) Ancestry informative markers and admixture proportions in northeastern Mexico. J Hum Genet 54:504-509.

Maynard Smith J and Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23-35.

McKeigue PM (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76:1-7.

Moltke I and Albrechtsen A (2014) RelateAdmix: A software tool for estimating relatedness between admixed individuals. Bioinformatics 30:1027-1028.

Morton NE (2003) Genetic epidemiology, genetic maps and positional cloning. Philos Trans R Soc B Biol Sci 358:1701-1708.

Norris ET, Wang L, Conley AB, Rishishwar L, Mariño-Ramírez L, Valderrama-Aguirre A and Jordan IK. (2018) Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics 19 Suppl 8:861

Patel SR, Celedon JC, Weiss ST and Palmer LJ (2003) Lack of reproducibility of linkage results in serially measured blood pressure data. BMC Genet 4 Suppl 1:S37.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E and Shriver MD (2001) Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198-207.

Purcell S, Neale B, Todd Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Debakker P, Daly MJ *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

Qin H and Zhu X (2012) Power comparison of admixture mapping and direct association analysis in genome-wide association studies. Genet Epidemiol 36:235-243.

Rife DC (1953) Fingerprints as criteria of ethnic relationship. Am J Hum Genet 5:389-399.

Risch N and Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Ros-Freixedes R, Battagin M, Johnsson M, Gorjanc G, Mileham AJ, Rounsley SD and Hickey JM (2018) Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. Genet Sel Evol 50:64.

Rubi-Castellanos R, Martínez-Cortés G, Muñoz-Valle JF, González-Martín A, Cerda-Flores RM, Anaya-Palafox M and Rangel-Villalobos H (2009) Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. Am J Phys Anthropol 139:284-294.

Salzano FM (2005) Evolutionary change - patterns and processes. An Acad Bras Cienc 77:627-650.

Salzano FM and Sans M (2014) Interethnic admixture and the evolution of Latin American populations. Genet Mol Biol 37:151-170.

Sans M (2000) Admixture studies in Latin America: From the 20th to the 21st century. Hum Biol 72:155-177.

Silva MCF, Zuccherato LW, Soares-Souza GB, Vieira ZM, Cabrera L, Herrera P, Balqui J, Romero C, Jahuira H, Gilman RH *et al.* (2010) Development of two multiplex mini-sequencing panels of ancestry informative SNPs for studies in Latin Americans: An application to populations of the state of Minas Gerais (Brazil). Genet Mol Res 9:2069-2085.

Skotte L, Jørsboe E, Korneliussen TS, Moltke I and Albrechtsen A (2019) Ancestry-specific association mapping in admixed populations. Genet Epidemiol 43:506-521.

Stumpf MPH and Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13:1-8.

Teng B, Yang C, Liu J, Cai Z and Wan X (2016) Exploring the genetic patterns of complex diseases via the integrative genome-wide approach. IEEE/ACM Trans Comput Biol Bioinforma 13:557-564.

The 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK *et al.* (2008) Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4:e4.

Tishkoff SA and Verrelli BC (2003a) Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340.

Tishkoff SA and Verrelli BC (2003b) Role of evolutionary history on haplotype block structure in the human genome: Implications for disease mapping. Curr Opin Genet Dev 13:569-575.

Winkler CA, Nelson GW and Smith MW (2010) Admixture mapping comes of age. Annu Rev Genomics Hum Genet 11:65-89.

Associate Editor: Jorge Lopez-Camelo

Capítulo 8

Anexo II

8.1. Artículo publicado en *Annals of Human Genetics*

Received: 11 February 2021 | Revised: 16 March 2021 | Accepted: 26 March 2021

DOI: 10.1111/ahg.12421

Annals of human genetics WILEY

SHORT COMMUNICATION

Population structure and relatedness estimates in a Mexican sample

V. Colistro[1] | A. Rojas-Martínez[2] | A. Carracedo[3,4] | CHIBCHA Consortium[*] | I. Tomlinson[5] | L. Carvajal-Carmona[6] | R. Cruz[4] | M. Sans[7]

[1] Departamento de Métodos Cuantitativos, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay

[2] Escuela de Medicina y Ciencias de la Salud, Tecnológico de Monterrey, Monterrey, México

[3] Centro Nacional de Genotipado (CEGEN), Universidad de Santiago de Compostela, Santiago de Compostela, Spain

[4] CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

[5] Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK

[6] Genome Center & Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, Davis, California, USA

[7] Departamento de Antropología Biológica, Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Montevideo, Uruguay

Correspondence
R Cruz, CIBER de Enfermedades Raras (CIBERER)-Instituto de Salud, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.
Email: raquel.cruz@usc.es
M. Sans, Departamento de Antropología Biológica, Facultad de Humanidades y Ciencias de la Educación, Universidad de la República, Montevideo, Uruguay.
Email: mbsans@gmail.com

[*]Members of CHIBCHA Consortium are listed in the Acknowledgments section.

Abstract

Population stratification (PS) is a confounding factor in genome-wide association studies (GWASs) and also an interesting process itself. Latin American populations have mixed genetic ancestry, which may account for PS. We have analyzed the relatedness, by means of the identity-by-descent (IBD) estimations, in a sample of 1805 individuals and 1.006.703 autosomal mutations from a case–control study of colorectal cancer in Mexico. When using the recommended protocol for quality control assessment, 402 should have been removed due to relatedness. Our purpose was to analyze this value in the context of an admixed population. For that aim, we reanalyzed the sample using two software designed for admixed populations, obtaining estimates of 110 and 70 related individuals to remove. The results showed that the first estimation of relatedness was an effect of the higher Native American contribution in part of the data samples, being a confounding factor for IBD estimations. We conclude in the importance of considering PS and genetic ancestry in order to avoid spurious results, not only in GWAS but also in relatedness analysis.

KEYWORDS

admixed populations and population stratification, identity-by-descent

1 | INTRODUCTION

Population stratification has been widely studied as a confounding factor in genome-wide association studies (GWASs). For that purpose, it is important to accurately identify individuals with a high probability of being related to a certain degree, otherwise, the association will be biased. Usually, the determination of identity-by-descent (IBD) is used to detect relatedness. For samples that came from structured populations, it is not valid to assume population homogeneity. In spite of dealing with presumably unrelated samples, identity-by-state (IBS) estimates

are essential in an early stage of any workflow for analyzing population structure.

Some authors have addressed the methodological issue of sameness due to shared ancestry as confounding in the determination of relatedness (Anderson & Weir, 2007; Wang, 2011). IBS is used in genetics to describe two identical alleles or two identical sequences of DNA, and these alleles may be identical by chance or inherited from a recent common ancestor. IBD is defined as the proportion of 0, 1, or 2 identical by descent alleles between two individuals; the higher the estimate, the greater the probability of relatedness. IBS represents the proportion of shared DNA segments, identical from the molecular point of view, but without sharing a common origin, or in which their common origin cannot be unequivocally determined (Forabosco et al., 2005). Nevertheless, hidden or unrecorded relations may cause bias in the estimates. Thornton et al. (2012), among others, have considered the problem of estimating relatedness in structured populations with admixed ancestry. However, having multiple source populations does not necessarily indicate population structure, in spite of multiple source populations can lead to more genetic diversity (Owings et al., 2019).

The purpose of this paper is to analyze the relatedness in a case–control study for colorectal cancer in the Mexican population (unpublished data), an example of admixed population, using different approaches. This concern arises from the first results for relatedness following the standard protocol (Anderson et al., 2010).

2 | MATERIALS AND METHODS

We analyzed 1805 samples from Mexico (929 cases and 876 controls), collected in three different locations: Mexico City, Monterrey, and Torreón. We have chosen Mexican population as an example of admixed population, as its European contribution ranges from 40 to 62% and the African from 1 to 6%, being the rest Native American (Salzano & Sans, 2014) . All participants provided written informed consent for inclusion. The study was conducted in accordance with the Declaration of Helsinki. The protocol was approved by the ethics committees of each participating institution and the Federal Commission for Protection against Health Risks (COFEPRIS, Mexico) (code CMN2012-001). Genotypes were obtained using two complementary arrays: Axiom Genome-Wide LAT 1 (Latino) Array and a Custom-designed Array, both from Affymetrix Axiom Genotyping Solutions. After SNP-level quality controls were applied, the resulting data set consisted of 1.006.703 autosomal SNPs uniformly distributed along the 22 autosomal chromosomes.

To assess the data quality of the analyzed samples, we carried out the stepwise protocol described in Anderson et al. (2010). This protocol deals with the quality control (QC) of genotype data from genome-wide association studies by using PLINK (Chang et al., 2015) to carry out assessments of failure rate per-individual and per-SNP and to determine the degree of relatedness between individuals. At some point, the protocol proposes a principal component analysis (PCA) for the identification of individuals of divergent ancestry. In that matter, we first pruned typed SNPs in high linkage disequilibrium (LD) with $r_2 > 0.1$ and then we determined the PCs using EIGENSTRAT within smartpca (Price et al., 2006).

To replicate the results of the IBD obtained with PLINK, we used two software specifically designed to estimate relatedness in structured populations. KING (Manichaikul et al., 2010) handles genotype data from GWASs or sequencing data and its algorithm considers the presence of population structure. Another program, REAP (Thornton et al., 2012), estimates autosomal kinship coefficients and IBD sharing probabilities using SNP genotype data in samples with admixed ancestry. REAP also requires to include allele frequencies of SNPs in the ancestral populations and admixture proportions of each individual in the sample. So, for these estimations, we used data available in the 1000 Genomes Project (The 1000Genomes Consortium, 2010). Samples were selected to represent the parental populations of Mexico assuming a trihybrid model: African (N = 176), European (N = 174), and Asian (N = 98); these last considered in substitution of Native Americans due to the scarcity of data about them, and the similarities because of the common origin. Estimation of individual admixture fraction was calculated with ADMIXTURE software version 1.3.1 (Alexander et al., 2009), which considers a likelihood model. To choose the correct value of k, we computed the cross-validation error over k from 2 to 6. We found that k_3 yielded the lowest cross-validation error ($k_3 = 0.538$).

3 | RESULTS

The estimation of IBD using PLINK showed an oddly high number of related samples, 402 samples to remove (23% of samples, IBD > 0.1875) (Table 1). This estimation is extremely unlikely since being a relative of another participant was an exclusion criterion during the sample collection process in Mexico. Most of the related samples formed blocks of individuals related all to all, rather than one to one (pairs) or one to a few. This made us think that some variable was generating false closeness. The samples reported as related were different when analyzing each array separately (92 in the CUSTOM array and 373

TABLE 1 Pairs of samples identified as related by three different software, according to the degree of relatedness

Relationship	PLINK	REAP	KING
Identical twins*	10	10	10
First degree	55	59	52
Second degree	24	32	29
Third degree	313	17	19
Total	402	73	110

*Repeated samples.

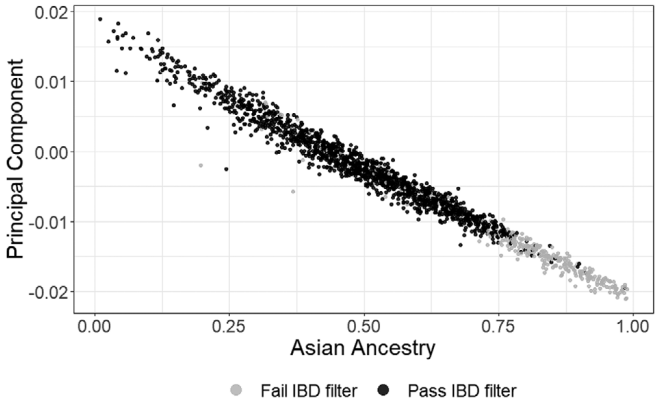

FIGURE 1 Principal component 1 against Native American ancestry. Pairs of samples with high IBD (>0.1875) are spotted in gray

in the Latino array). To understand which factors were being confounded with relatedness, we evaluated several variables: sampling location, case/control status, ancestry estimation, sex, genotype batch, and significant principal components. The admixture analysis stratified by case/controls showed no significant differences between groups. However, cases tend to present greater Caucasian ancestry (41.27% vs. 38.5% in controls), while controls presented an increased Asian ancestry (55.9% vs. 53.23%). The African ancestry was the smallest in both groups, 5.6% in cases and 5.02% in controls. The mean global ancestry in the whole dataset was 54.3% Asian, 40.4% of Caucasian, and 5.3% of African. We detected that Asian ancestry was highly correlated PC1, and pairs of samples with high IBD positioned at the very end of the distribution (higher Asian ancestry and lower eigenvalue for PC1) (Figure 1).

Lately, we assumed structured samples and used KING and REAP software. Both consider the structure of the samples due to shared ancestry to avoid identifying samples of the same ancestry as related. When running KING, 110 samples have to be removed (Table 1). The results of REAP showed consistency with KING, but in this case, only 73 samples to remove were identified (Table 1). To assure comparability, we used the IBD thresholds proposed in KING to determine the degree of relatedness on REAP results.

4 | DISCUSSION

Genetic diversity is lower in Native American populations related to other continental regions, measured by heterozygosity (Wang et al., 2007). On the other hand, Mexicans have a high degree of differentiation between populations explained by high degrees of isolation, measured by FST that quantifies the proportion of the genetic variance contained in a subpopulation relative to the total genetic variance. Moreover, those populations appear as discrete units, with scarce gene flow (Moreno-Estrada et al., 2014). We observed that higher IBD values are related to increased Native American ancestry when using PLINK. We postulate that the isolation that leads to a high differentiation among Mexican Natives is a confounding factor for IBD estimation due to deep sub-structuration, and each ethnic group can be taken as a familiar group as happens with software not developed to deal with admixed samples. With Mexican mestizo samples, we have shown that the software confounded Native American ancestry with relatedness, overestimating the relatedness among samples. This has practical implications when studying admixed populations as we have shown that exclusion of samples based on measures of relatedness may produce biased results. Lastly, differences found between both arrays can be explained because the Latino Array was designed to highlight differences among parental populations.

ACKNOWLEDGMENTS

CHIBCHA Consortium (study of hereditary cancer in Europe and Latin America) members are as follows: Ian Tomlinson (University of Edinburgh, UK); Luis Carvalal-Carmona (University of California, Davis, USA), Ma. Magdalena Echeverry de Polanco, Mabel Elena Bohórquez, Rodrigo Prieto, Angel Criollo, Carolina Ramírez, Ana Patricia Estrada, Jhon Jairo Suárez (Universidad del Tolima, Colombia); Augusto Rojas Martinez, Rocío Ortiz Lopez (Tecnológico de Monterrey, Mexico); Silvia Rogatto, Samuel Aguiar Jnr, Ericka Maria Monteiro Santos (São Paulo State University, Botucatu, Brazil); Monica Sans, Valentina Colistro, Pedro C. Hidalgo, Patricia Mut (University of the Republic, Uruguay); Angel Carracedo, Clara Ruiz Ponte, Ines Quntela Garcia (University of Santiago de Compostela, Spain); Sergi Castellvi-Bel (University of Barcelona, Barcelona, Catalonia, Spain); Manuel Teixeira (Portuguese Oncology Institute, Portugal).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The dataset generated and analyzed during the current study are available from corresponding authors on reasonable request.

AUTHOR CONTRIBUTIONS

Study design: L. Carvajal-Carmona and I. Tomlinson; conceptualization: V. Colistro, A. Rojas-Martínez, I. Tomlinson, L. Carvajal-Carmona, and M. Sans; data curation: R. Cruz; formal analysis: V. Colistro, A. Carracedo, and R. Cruz; methodology: V. Colistro, R. Cruz, and M. Sans; funding acquisition: A. Rojas-Martínez, A. Carracedo, CHIBCHA Consortium, I. Tomlinson, and L. Carvajal-Carmona; visualization: V. Colistro; project administration: A. Rojas-Martínez, A. Carracedo, CHIBCHA Consortium, I. Tomlinson, and L. Carvajal-Carmona; supervision: M. Sans; writing original draft: V. Colistro, A. Rojas-Martínez, and M. Sans.

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, *176*(1), 421–440. https://doi.org/10.1534/genetics.106.063149

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, *5*(9), 1564–1573. https://doi.org/10.1038/nprot.2010.116.Data

Chang, C. C., Chow, C. C., Tellier, L. C. A. M. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Forabosco, P., Falchi, M., & Devoto, M. (2005). Statistical tools for linkage analysis and genetic association studies. *Expert Review of Molecular Diagnostics*, *5*, 781–796. https://doi.org/10.1586/14737159.5.5.781

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*, *344*(6189), 1280–1285. https://doi.org/10.1126/science.1251688

Owings, A. C., Fernandes, S. B., Olatoye, M. O., Fogleman, A. J., Zahnd, W. E., Jenkins, W. D., Malhi, R. S., & Lipka, A. E. (2019). Population structure analyses provide insight into the source populations underlying rural isolated communities in Illinois. *Human Biology*, *91*(1), 31–47. https://doi.org/10.13110/humanbiology.91.1.05

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. a, & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Salzano, F. M., & Sans, M. (2014).. Interethnic admixture and the evolution of Latin American populations. *Genetics and Molecular Biology*, *37*(1), 151–170. https://doi.org/10.1590/S1415-47572014000200003

The 1000Genomes Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *American Journal of Human Genetics*, *91*(1), 122–138. https://doi.org/10.1016/j.ajhg.2012.05.024

Wang, J. (2011).. Unbiased relatedness estimation in structured populations. *Genetics*, *187*(3), 887–901. https://doi.org/10.1534/genetics.110.124438

Wang, S., Lewis, C. M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M. V., Molina, J. A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A. M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M. C., Salzano, F. M., …, & Ruiz-Linares, A. (2007).. Genetic variation and population structure in Native Americans. *PLoS Genetics*, *3*(11), e185. https://doi.org/10.1371/journal.pgen.0030185

How to cite this article: Colistro V, Rojas-Martínez A, Carracedo A, et al. Population structure and relatedness estimates in a Mexican sample. *Ann Hum Genet*. 2021;1–4. https://doi.org/10.1111/ahg.12421

Serpiente Emplumada