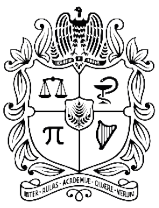


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Perfiles moleculares del cáncer colo-rectal en diferentes poblaciones Colombianas

María Carolina Sanabria Salas

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Química
Bogotá, Colombia
2017

Perfiles moleculares del cáncer colo-rectal en diferentes poblaciones Colombianas

MD., MSc., María Carolina Sanabria Salas

Tesis de investigación presentada como requisito parcial para optar al título de:

Doctorado en Ciencias - Química

Director:

MSc., Ph.D., Adriana Umaña Pérez

Codirector:

MD., MSc., Gustavo Hernández Suárez

Línea de Investigación en Bioquímica

Grupo de Investigación en Hormonas

Universidad Nacional de Colombia

Facultad de Ciencias, Departamento de Química

Bogotá, Colombia

2017

Dedico este trabajo a mis padres y a mis hermanos por su apoyo durante este proceso, a Dios y muy especialmente a mi Angelita de la guarda.

Agradecimientos

A la Universidad Nacional de Colombia, UNAL, por permitirme crecer académicamente durante mis estudios de doctorado.

Al Instituto Nacional de Cancerología, INC, especialmente a quienes han sido sus Directores Generales en los últimos años, Carlos Vicente Rada, MD., Raúl Murillo, MD., MSc., y Carolina Wiesner MD., MSc., por haberme dado la oportunidad de desempeñarme laboralmente dentro de la Institución y, sobretodo, por permitirme cursar mis estudios de doctorado a la par. Igualmente, por el apoyo financiero a los proyectos de cáncer colorrectal en genética (C41030310-24) y en proteómica (C41030610-119).

A Gustavo Hernández, MD., MSc., por haberme recibido como estudiante de doctorado en este proyecto tan bonito e innovador en Colombia y asesorarme como codirector de tesis en el proceso, especialmente en el análisis de datos genéticos.

A Martha Serrano, MD, MSc, PhD., por haber confiado en mí como médico genetista para trabajar en este proyecto de doctorado y recomendarme profesionalmente en el cargo en el cual me desempeño actualmente en el INC.

A Alba Lucía Combita, PhD., por su apoyo tanto a nivel personal como profesional y académico, y a todos mis demás compañeros de trabajo en el Grupo de Investigación en Biología del Cáncer por su apoyo.

A la profesora Myriam Sánchez de Gómez, MSc, por aceptarme como integrante del grupo de Investigación en Hormonas de la UNAL, dentro del cual pude realizar mis actividades académicas.

A los integrantes del grupo de Investigación en Hormonas, por aportarme conocimientos en bioquímica y vías de señalización en cáncer, permitiéndome adquirir un entendimiento integral de mí de investigación. Especialmente, a Ruth Andrea Rodríguez, quién me tuvo mucha paciencia con todos los temas químicos

A la Dirección de Investigación Sede Bogotá, por el apoyo financiero a los proyectos en genética y proteómica del cáncer colorrectal (DIB 20708).

A Colciencias por financiar el proyecto de genética en cáncer colorrectal (210145921617)

A Jovanny Zabaleta, PhD., director del Genomics Core de la Universidad de Louisiana, por asesorarme en la realización de técnicas de genotipificación a gran escala.

A Brett Phinney, PhD., director del Proteomics Core de la Universidad de California en Davis y a Darren Weber, investigador asociado, quiénes me acogieron en el core y me asesoraron en la preparación de las muestras para los análisis en HPLC/MS-MS y en el análisis de los espectros de masas obtenidos.

Al profesor Jairo Arturo Guerrero, de la cátedra Técnicas Cromatográficas Modernas de la UNAL, con quién aprendí mucho de cromatografía y espectrometría de masas, y gracias a quién le tome cariño a estos temas tan nuevos para mí.

A Albert Tenesa, PhD., quién me recibió en su grupo de análisis bioinformático en el Instituto Roslin de la Universidad de Edimburgo y a Konrad Rawlik, PhD., quién trabajo conmigo de cerca para avanzar en los análisis de ancestría genética local.

A Laura Fejerman, PhD., por sus aportes importantísimos en los análisis genéticos del estudio y su apoyo en la escritura del artículo.

A mi directora de tesis, la profesora Adriana Umaña, PhD., por aceptarme como su estudiante y tenerme mucha paciencia durante este largo proceso, así como por su orientación especialmente con el componente proteómico.

Resumen

La incidencia y mortalidad por cáncer colorrectal (CCR) son variables entre las diferentes regiones de Colombia; esto puede estar influenciado por factores ambientales y genéticos, incluyendo la ancestría, que es una mezcla entre europeos, africanos y amerindios. El CCR en estadios tempranos es curable hasta en el 90% de los pacientes, sin embargo, es diagnosticada principalmente en estadios avanzados cuando el pronóstico es pobre; esto se debe a que no existen pruebas con alta sensibilidad y especificidad, y que sean mínimamente invasivos, para el diagnóstico temprano de la enfermedad y su seguimiento. La búsqueda de biomarcadores de susceptibilidad, diagnóstico y pronóstico, es una necesidad para mejorar el control del CCR en nuestro país. Entonces, planteamos buscar biomarcadores de susceptibilidad para tumores colorrectales, mediante estudios a mediana/gran escala de variantes genéticas comunes en una muestra multiregional de casos y controles de Colombia, incluyendo dentro de los análisis las proporciones de ancestría genética. Adicionalmente, planteamos la búsqueda de perfiles proteómicos diferenciales en colombianos con CCR usando técnicas de alta resolución en proteómica, con el fin de proponer biomarcadores de diagnóstico o pronóstico. Al respecto, encontramos diferencias importantes en la estructura genética de las poblaciones colombianas, en acuerdo con nuestra historia desde la colonización, y que las proporciones de ancestría se asociaron al riesgo de tumores colorrectales. Finalmente, proponemos dos variantes genéticas de riesgo para tumores colorrectales y un perfil de proteínas en plasma, como posibles biomarcadores de susceptibilidad y de pronóstico en CCR, respectivamente, que podrían implementarse para su control en Colombia.

Palabras clave: Análisis de Secuencia por Matrices de Oligonucleótidos; Proteómica; Neoplasias Colorrectales; Biomarcadores

Abstract

Colorectal cancer (CRC) incidence and mortality rates vary among the different regions of Colombia; this may be influenced by environmental and genetic factors, including ancestry, which is a mixture between Europeans, Africans and Amerindians. CRC at early stages is curable in almost 90% of the cases, however, it is mainly diagnosed in advanced stages when the prognosis is poor; this is explained because of the lack of minimally invasive tests with high sensitivity and specificity for early detection of the disease and its follow-up. The search for biomarkers of susceptibility, diagnosis and prognosis is a necessity to improve the control of CRC in our country. We then propose to search for biomarkers of susceptibility for colorectal tumors, through the implementation of medium/large scale studies of common genetic variants in a multiregional sample of cases and controls from Colombia; we included within the analyzes the proportions of genetic ancestry. In addition, we propose the search for differential proteomic profiles in Colombians with CRC using high resolution techniques in proteomics, in order to propose diagnostic or prognostic biomarkers. In this regard, we find important differences in the genetic structure of Colombian populations, in agreement with our history since colonization and, also, that the ancestry proportions were associated with the risk of colorectal tumors. Finally, we propose two genetic variants of risk for colorectal tumors and a plasma protein profile as possible biomarkers of susceptibility and prognosis in CRC, respectively, which could be implemented for its control in Colombia.

Keywords: DNA Microarrays; Proteomics; Colorectal Neoplasms; Biomarkers

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Lista de Abreviaturas y Glosario

Abreviatura	Término
1000 Genomes	1000 genomas, base de datos
2DE	Electroforesis bidimensional en geles de poliacrilamida
A1AG1	Alpha-1-acid glycoprotein 1
A1AT	Alpha-1 antitrypsin
ACN	Acetonitrilo
ADMIXTURE	Fast ancestry estimation tool, programa bioinformático
aDNA	DNA antiguo
AFR	Ancestría africana
AGC	Automatic Gain Control. Parámetro para limitar el número de iones que entran en la trampa.
AIC	Criterio de información de Akaike
AINES	Antiinflamatorios no esteroideos
AMBIC	Bicarbonato de amonio
AME	1. Poblaciones amerindias Pima, Maya, Karitiana, Surui y de Colombia. 2. Ancestría amerindia
Ant_F	Historia familiar de cáncer colorrectal
APC	Adenomatous polyposis coli, gen
APOH	Apolipoproteína H
Asn	Asparagina
B1R	Receptor inducible 1 de kininas
B2R,	Receptor constitutivo 2 de kininas
B4E1F0	Inhibidor de la proteasa plasmática C1
B-catenina	Proteína importante en la vía Wnt estimulando la expresión de genes blanco al actuar junto con factores de transcripción TCF7L2
Bcl-2	Familia de proteínas anti-apoptóticas
BMP	Proteína morfogénica ósea, factor de crecimiento
BMPR1	Bone morphogenetic protein receptor, type IA, gen
C1S	Componente de la proteína de complemento C1
CA19-9	Antígeno carbohidrato 19-9, marcador de pronóstico del cáncer colorrectal
CCR	Cáncer colorrectal
CD14	Antígeno de diferenciación de monocitos

Abreviatura	Término
CD169	Molécula de adhesión y de interacción en macrófagos
CDX2	Proteína homebox, factor de transcripción expresado por células epiteliales intestinales. Marcador de diferenciación intestinal
CEA	Antígeno carcino-embrionario, marcador de pronóstico del cáncer colorrectal
CEU	Residentes de Utah con ancestría europea
CFAB	Proteína del sistema de complemento, Complement factor B
CG	Genes candidatos, microarreglo
CHB	Asiáticos de Beijing, China
CI	Intervalo de confianza
Ciclina D1	Proteína que promueve proliferación celular
Ciu	Ciudad de origen
COL	Colombianos a estudio
COX-2	Ciclooxigenasa 2
C-trap	Trampa de iones en forma e C (parte del equipo Q-Exactive Orbitrap)
DANE	Departamento Administrativo Nacional de Estadística de Colombia
Df	Grados de libertad
DNA	Ácido desoxiribonucléico
DTT	Ditiotreitol
ECM	Matriz extracelular
EDTA	Ácido EtilenDiaminoTetraAcético
Edu	Nivel educativo
EM	Algoritmo esperanza-maximización
ER	Receptor de estrógenos
ESI	Método de ionización por electrospray
EUR	Ancestría europea
FA	Ácido fórmico
FAP	Poliposis Adenomatosa Familiar
FBLN1	Gen que codifica la proteína fibulina 1
FBNL1	Fibulina 1
FDA	US Food and Drug Administration,
FDR	False Discovery Rate
FGF-2	Factor de crecimiento de fibroblastos 2
FGFs	Factor de crecimiento de fibroblastos
FIBA	Cadena alfa del fibrinógeno
GLM	Regresiones lineales generalizadas
GLOBOCAN	Estimaciones de incidencia, mortalidad y prevalencia de los principales tipos de cáncer, base de datos
GREM1	Gen que codifica la proteína gremlin 1, que es inhibidor de la vía del TGF beta
GWAS	Estudios de asociación de genoma completo, microarreglo
HA	Ácido hialurónico
HapMap3 project	Mapa de haplotipos del genoma humano, base de datos
HCD	Colisión de alta energía para fragmentar iones y secuenciar los péptidos
HeatMap	Tipo de análisis para el agrupamiento jerárquico de grupos según sus características (datos)

Abreviatura	Término
Hedghogs	Familia de proteínas que participan en la diferenciación celular
Het	Heterocigoto
HGDP	Human Genome Diversity Cell Line Panel, base de datos
HMWK	Kininógenos de alto peso molecular
HNPCC	Cáncer colorrectal hereditario no polipósico
Hom	Homocigoto
HPLC	Cromatografía liquida de alta eficiencia
HPLC-MS (MS-MS)	Sistema de separación cromatográfica acoplada a un espectrometro de masas (en tándem)
hTERC	Gen que codifica el componente RNA de la telomerasa
HUPO	Organización del proteoma humano
HWE	Equilibrio de Hardy-Weinberg
IAA	Iodoacetamida
IBD	Identidad por descendencia
IBS	Población de España
IBS	Identidad en estado
Ig	Inmunoglobulinas
IgG	Inmunoglobulina G
IMC	Índice de masa corporal
Indels	Inserciones/deleciones
IPG	Gradiente de pH inmóvil
ITI	Familia de inhibidores de proteasas
ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3
ITIH4	Inter-alpha-trypsin inhibitor heavy chain H4
k	Número de poblaciones ancestrales asumidas en las estimaciones de ancestría de una muestra con mezcla racial
KLKB1	Plasma kallikrein
LAI	Inferencia de ancestría local
LD	Desequilibrio de ligamiento
LLA	Leucemia linfoblástica aguda
Log2()	Logaritmo en base 2, tipo de transformación de datos
logit()	Transformación logit de los datos
LV401	Región variable de la cadena ligera lambda de inmunoglobulinas
LWK	Africanos de Webuye, Kenia
Lys	Lisina
m/z	masa/carga
MAF	Frecuencia del alelo menor
MAPK	Proteína quinasa activada por mitógenos
MD2	Antígeno de linfocitos 96
MDS	Análisis de escalamiento multidimensional
MEX	Residentes de California descendientes de mexicanos
miRNA	Micro RNA

Abreviatura	Término
MMPs	Metaloproteasas de matriz
MMR	Mismatch repair, genes
mRNA	RNA mensajero
MSI	Inestabilidad de microsatétites
MS	Espectrometría de masas
MS-MS	Espectrometría de masas en tándem
mtDNA	DNA mitocondrial
MYC	Familia de protooncogenes blanco de la vía Wnt/B-catenina
MYH	MutY DNA glycosylase, gen
n	Número de individuos analizados o que hacen parte de un grupo
NF-κβ	Complejo proteico que controla la transcripción del DNA
OR	Odds Ratio
P	*P* valor
PA	Pólipos Adenomatosos
PCA	Análisis de componentes principales, usado para reducir la dimensionalidad de un set de datos
PCNA	Antígeno nuclear de celulas proliferantes
PCs	Componentes principales. Obtenidos a partir de un análisis tipo PCA
PGE2	Prostaglandina E_2
pI	Punto isoeléctrico
PLINK	Whole genome association analysis toolset, programa bioinformático
PR	Receptor de progesterona
PSA	Antígeno prostatico sérico
PTEN	Phosphatase and Tensin Homolog, gen
Q-Exactive Orbitrap	Equipo de espectrometría de masas que consiste en un cuadrupolo para filtrar el grupo de iones a separar y fragmentar en el analizador Orbitrap
R statistics	Programa de análisis estadístico
REM	Razones de mortalidad estandarizadas
RFMix	A discriminative method for local ancestry inference, programa bioinformático
RNA	Ácido ribonucléico
RPLC	Cromatografía líquida en fase reversa
RPLC/MS-MS	Cromatografía en fase reversa acoplada a espectroetría de masas en tándem
Scaffold	Programa bioinformatico para visualizar, analizar y validar resultados proteómicos
SDS-PAGE	Electroforesis en gel de poliacrilamida en condiciones denaturantes
SE	Error estándar
SERPINAS	Superfamilia de proteínas que inhiben a serina proteasas
ShapeIT	A linear complexity phasing method for thousands of genomes, programa bioinformático

Abreviatura	Término
SMAD4	Mothers against decapentaplegic homolog 4, gen
SNPs	Polimorfismos de un solo nucleótido
STK11	Serine/threonine kinase 11, gen
TAE	Tasas ajustadas por edad
tagSNP	SNP que permite rastrear a otro (con efecto funcional), al estar fisicamente ligados o cercanos
TaqMan	Ensayo de genotipificación
TCF7L2	Factor de transcripción, efector de la vía de señalización Wnt/B-catenina
TEP1	Gen que codifica a la proteína asociada a la telomerasa 1
TFA	Ácido trifluoroacético
TGF beta	Factor de crecimiento transformante beta
TIMs	Macrófagos infiltrantes de tumor
TK1	Gen que codifica a la proteína timidina kinasa 1
TLR	Receptor de lipopolisacáridos
TP53	Tumor protein p53, gen
tPA	Activadores de plasminógeno tisular
Tris	Tris(hidroximetil)aminometano
TROVE	Dominio de unión al RNA con funcion regulatoria en la elongación de los telóme
uPA	Activadores de plasminógeno urokinasa
USA (US)	Estados Unidos de América
Wnt	Grupo de proteínas importantes en señalización celular: motilidad, polaridad, organogénesis y renovación de células madre
X! Tandem	Programa bionformático para comparar espectros de masas con secuencias de péptidos para la identificación de proteínas
YRI	Africanos de Ibadan, Nigeria
ZA2G	Zinc-alpha-2-glycoprotein
λ	Factor de inflación lambda

Concepto	Definición
Alelos	Copias o formas de un gen.
Algoritmo esperanza-maximización (EM)	Método probabilístico para estimar la máxima verosimilitud en modelos con datos incompletos. El objetivo de éste método es encontrar los parámetros que maximizan la probabilidad de los datos observados.
Análisis componentes principales (PCA)	Método para reducir la dimensionalidad de un conjunto de datos que pueden estas correlacionados, con el fin de obtener un nuevo conjunto de valores de variables sin correlación lineal llamadas componentes principales (PCs). Lo anterior, permite resumir la varianza de los datos originales en los PCs, en donde el primer PC es el que capta la mayor proporción de la varianza, seguido por el segundo PC, y así sucesivamente.
Ancestría genética	Concepto genético que describe la arquitectura de la variación genética entre poblaciones; por lo tanto, calcular la ancestría genética a nivel poblacional e individual permite medir de manera objetiva que proporción de la variabilidad genética es compartida con otros grupos étnicos o raciales. Esta medición, es especialmente importante en poblaciones con alto grado de mezcla, con el fin de corregir los análisis de asociación a nivel poblacional.
Aneuploidía	Término usado para referirse a la ganancia o pérdida de cromosomas; es decir, consiste en cambios en el número de cromosomas en las células.
Aptitud reproductiva	Describe el éxito reproductivo de un individuo. Equivale a la contribución promedio de individuos con un genotipo o fenotipo característico al pool genético de la siguiente generación.
Arquitectura alélica	Se refiere al número de alelos que impactan en el riesgo de una enfermedad, la penetrancia de éstas variantes y la frecuencia de estas en la población a estudio. Está en relación con el concepto de heredabilidad, pues con ambos se busca entender que tanto de la variabilidad fenotipica se explica por la variabilidad genética.
Autosomas	Son los pares de cromosomas del 1 al 22 (no incluye los sexuales X y Y). Cada persona hereda información de los autosomas tanto de su padre como de su madre, por lo tanto, se denominan biparentales. A nivel poblacional, las mujeres aportan la mitad de todos los autosomas en la población (1/2) y los hombres aportan la otra mitad (1/2). Debido a que en cada persona ocurren eventos de recombinación genética entre los pares homólogos de autosomas, el estudio de éstos ha ha permitido entender mejor diferentes conceptos relacionados con la genética de poblaciones, como la migración, entre otros.
Blancos terapéuticos	Son moléculas o vías de señalización sobre el cual un fármaco actúa de manera específica.
Células mesenquimales	Son células madre estromales; se caracterizan por ser multipotenciales primitivas que se originan a partir de la capa germinal mesodermal del embrion. Son altamente proliferativas y tienen la capacidad de diferenciarse en diversos tipos de células.
Disociación por colisión de alta energía (HCD)	Consiste en una técnica usada en MS para fragmentar iones moleculares en fase gaseosa. Los iones son acelerados por potenciales eléctricos proporcionándoles una alta energía cinética y favoreciendo la colisión de éstos con moléculas de un gas neutro como helio, causando su fragmentación. Este proceso es necesario para secuenciar los péptidos.

Concepto	Definición
Criptas intestinales	Son invaginaciones del epitelio superficial intestinal y se dividen en dos zonas: *la zona proliferativa* o nicho de células madre que se ubica en la base, y *la zona de diferenciación* que se ubica hacia la luz intestinal. Allí ocurre recambio celular cada tres a seis días para generar nuevas células madre que se diferenciarán en células epiteliales intestinales y así mantener el tejido. Un desequilibrio en el patrón normal de recambio, por alteración en las vías de señalización implicadas, favorece la persistencia de señales proliferativas y antiapoptóticas, lo que lleva a un agrandamiento y fisión de las criptas, formando focos de criptas aberrantes que son características en los tejidos tumorales colorrectales.
Criterio de información de Akaike (AIC)	Es un método para seleccionar el mejor modelo al ser una medida de la calidad relativa del mismo. El modelo es el del mínimo valor. Su fórmula incluye la probabilidad de los datos según el modelo usado y el número de parámetros o variables usados en el modelo. Este último es penalizado con el fin de disminuir el sobreajuste de los modelos.
Cromosoma X	Cromosoma sexual que hace parte del par 23 en seres humanos. Las mujeres heredan un X de su padre y el otro de madre (XX; copias homólogas = marcadores homocigotos o heterocigotos). El hombre solo puede heredar un X de su madre y el Y de su padre (XY; una copia = marcadores hemicigotos). A nivel poblacional, existe un total de copias de X relativo de tres; las mujeres aportan dos copias del cromosoma X (2/3) y los hombres una copia (1/3). Vale la pena aclarar que el cromosoma X de los hombres es el reflejo de los cromosomas X de las mujeres en la generación anterior, pues lo heredaron de su madre. También es importante resaltar que existen regiones recombinantes y no recombinantes a lo largo de este cromosoma; los marcadores en la región no recombimante se asignan al cromosoma 23, mientras que los marcadores en la región recominante o pseudo-autosómica se asignan al *"cromosoma 25"* que no existe físicamente, pero se asigna así para efectos de la manipulación de las bases de datos genéticas disponibles.
Cromosoma Y	Cromosoma sexual (XY; par 23 en seres humanos). Solo puede ser heredado del padre, por lo que se considera uniparental. Tiene una región no recombinante y una región pseudo-autosómica con la cuál existe una mínima recombinación con el cromosoma X. En genética de poblaciones, los marcadores en el cromosoma Y hablan del linaje paterno en el estudio de eventos de migración, apareamiento y mezcla genética en el surgimiento de poblaciones más recientes.
Deriva génica	Fuerza evolutiva que junto con la selección natural, cambia las frecuencias de los alélos de una especie con el tiempo. Ocurre por efecto de la reproducción y la pérdida de alelos al azar; generalmente se pierden los alelos menos frecuentes, mientras que favorece la permanencia o fijación de los más frecuentes. Por lo anterior, un efecto importante, es que la población se vuelve mas homogénea al disminuir la diversidad genética. Estos efectos son más acentuados en las poblaciónes pequeñas en número de habitantes, las cuales tienen a volverse cada vez más homocigotas. Un ejemplo de esto, es lo que ocurre en poblaciones con efecto fundador.

Concepto	Definición
Desequilibrio de ligamiento (LD)	Propiedad de algunos marcadores genéticos de segregarse en conjunto debido a su cercanía física. Entre mas cercanos esten dos marcadores, menos probable es que ocurra recombinación entre ellos. Solo cuando hay recombinación entre dos marcadores, estos pueden segregarse de forma independiente en la siguiente generación. Marcadores en el mismo cromosoma con una frecuencia de recombinación menor del 50%, se considera que pueden estar en LD. Existen dos medidas del grado de LD entre dos marcadores, el D' y el r^2. Si D' = 1 se considera que los dos marcadores están en completo LD, si D' = 0 se considera que estos son totalmente independientes y que no se segregan juntos y si D' = 0.5, quiere decir que el 50% de las veces estos marcadores se segregan juntos. Por otro lado, r^2 no solamente habla de la correlación entre dos marcadores como medida de que se segregan juntos, sino que también tiene en cuenta la frecuencia alélica de estos. Por lo tanto, aún cuando dos marcadores estén ligados, es posible que si la frecuencia alélica de uno de ellos es muy rara, no sea posible ver su correlación tan frecuente. Entonces, un r^2 = 1 indica que los dos marcadores están en perfecto LD, mientras que valores menores se deben evaluar también en relación al D' para su interpretación.
Diploidía	Dos juegos completos de cromosomas en el núcleo de una célula. En humanos una célula diploide tiene 23 pares de cromosomas. Habla de la estabilidad cromosómica de la célula.
Disparidad	Desigualdad o diferencias. En salud pública se usa este término para referirse a las implicaciones de las diferencias o desigualdades en la variación del estado de salud de los individuos en una población o entre poblaciones. Causas de disparidad en salud a nivel poblacional pueden ser por ejemplo: el nivel socio-económico, el acceso a la educación, el aceso a la salud, diferencias biológicas o genéticas, diferencias en el sexo, entre otros.
DNA antiguo (aDNA)	Se refiere al DNA aislado de especímenes antíguos, lo que implica la recuperación de material biológico de muestras que no fueron preservadas para tal fin. El análisis de este tipo de material ha permitido avanzar en el estudio de la evolución de diferentes especies, lo cual ha sido importante en el entendimiento de la genética de las poblaciones.
DNA mitocondrial (mtDNA)	Material genético almacenado en las mitocondrias de las células. Este se reproduce por sí mismo con cada división celular. Los espematozoides contienen unos pocos cientos de moléculas de DNAmt por célula en la cola; mientras que los oocitos tienen hasta 200.000 copias. Durante la fecundación, el espermatozoide pierde la cola, por lo tanto, en genética de poblaciones, el análisis de DNA mitocondrial habla del linaje materno (marcadores uniparentales) en el estudio de eventos de migración, apareamiento y mezcla genética en el surgimiento de poblaciones más recientes.
Enhancer	Un enhancer es un potenciador o amplificador que actúa aumentando los niveles de la transcripción de genes. Consiste en una secuencia corta del DNA eucariota (de 50 a 1500 pares de bases) que puede unirse con proteínas con función de factores de transcripción y con la enzima polimerasa II para aumentar la expresión génica. Actúan en cis sobre genes ubicados hasta 1 millón de pares de bases de su localización. Se han descrito cientos de miles de enhancers a lo largo del genoma humano.
Equilibrio de Hardy-Weinberg (HWE)	Es un modelo que describe la relación entre las frecuencias alélicas y genotípicas en una población diploide sexualmente reproductiva con apareamiento al azar. Se considera que un marcador está en desequilibrio de HW cuando las frecuencias genotípicas observadas difieren significativamente de las esperadas, asumiendo una población en equilibrio, en la cual no ha ocurrido selección, ni mutaciones, ni migraciones y es lo suficientemente grande como para descartar efectos por deriva génica.

Concepto	Definición
Escalamiento multidimensional (MDS)	Consiste en un método para detectar las dimensiones que mejor explican las similitudes o diferencias, es decir las distancias, entre objetos. Permite observar gráficamente las distancias entre los objetos de estudio, usualmente en dos dimensiones, por lo que se considera que es un metodo para reducir la complejidad de los datos a analizar.
Especificidad	Es la probabilidad de que un sujeto sano tenga un resultado negativo en una prueba, por ejemplo de diagnóstico. Es decir, es el porcentaje de verdaderos negativos (sanos), por lo que permite saber la capacidad de una prueba para descartar a todos los que son sanos libres de la enfermedad a estudio.
Factor de inflación lambda	En estudios de asociación genética, este es una parámetro usado para obtener evidencia de que existen diferencias sistemáticas en el estudio que pueden actuar como confusores y aumentar de esta manera la obtención de falsos positivos en los análisis de asociación. Algunos confusores pueden ser: estratificación poblacional, parentesco entre las muestras o errores en el genotipado no detectados en los pasos de control de calidad por individuo y por SNPs; todos estos pueden ser los responsables de las diferencias observadas en las frecuencias alélicas, más no el estatus de enfermedad, que es lo que se está evaluando. Por lo anterior, es importante identificar si existen confusores con el fin de corregir los análisis de asociación. Un lambda ~ 1.00 indica que no existe inflación en los resultados obtenidos, que es lo que se busca.
Factores confusores	Son aquellas variables que se correlacionan tanto con la variable dependiente como con la independiente. Su presencia crea sesgos en los análisis y sugerir asociaciones que no son reales. Es importante identificar potenciales confusores, con el fin de medirlos y usarlos para corregir los análisis.
Falla en el genotipado	Tasa de no genotipado de un marcador, lo que puede deberse a problemas relacionados con la sonda usada para amplificar la secuencia de interés o por interferencias en la muestra usada.
Faseado de genotipos en haplotipos	Se refiere a determinar que grupo de alelos estan ligados o se heredaron de un mismo progenitor. Si tenemos un par de genotipos Aa y Bb para los marcadores 1 y 2, respectivamente, en un individuo; entonces, fasear esta información es modelar si este individuo heredo la combinación AB o Ab o aB o ab en bloque de uno de sus progenitores, heredando del otro las posibles combinaciones ab, aB, Ab o AB, respectivamente. De esta manera se arman los haplotipos a partir de los genotipos permitiendo rastrear así la información heredada de nuestros ancestros.
Fenotipo mutador	Característica que resulta de la ganancia de múltiples mutaciones genéticas debido al funcionamiento anormal de proteínas reparadoras de DNA. Mutaciones o el silenciamiento de genes que codifican proteínas de reparación del DNA por mal apareamiento, generan una falla en el sistema de reparación del DNA y el acúmulo de mutaciones en multiples regiones del genoma, causando inestabilidad genética y mayor riesgo de cáncer.
Frecuencia del alelo menor (MAF)	Es la frecuencia del alelo menos frecuente de un SNPs en una población. Este dato permite diferenciar entre variantes raras o comunes en una población, lo cual es muy útil en genética de poblaciones y en el estudio de enfermedades complejas. El proyecto del genoma humano HapMap, incluyó los genotipos de multiples SNPs con un MAF mayor o igual al 5% en diversas poblaciones, con el fin de obtener evidencia sobre la variabilidad genética entre éstas.

Concepto	Definición
Fundadores	En términos poblacionales se refiere a personas que dieron origen a las siguientes generaciones en una población.
Genes supresores de tumor	Gen que codifica una proteína que actúa inhibiendo vías relacionadas con la proliferación celular, mientras que favorecen las vías relacionadas con la apoptosis. Lo anterior,con el fin de contrarrestar el efecto de los protooncogenes y ejercer control en el ciclo celular. Se requiere la ganancia de alteraciones genéticas (como mutaciones o deleciones, entre otros, o la combinación de varios) en las dos copias de este gen, para inactivarlo.
Genotipos (genotipado)	Genotipos se refiere a la combinación de alelos para un marcador, en el cual uno fue heredado del padre y el otro de la madre. Y genotipado se refiere a la acción de realizar técnicas moleculares para la obtención del genotipo de interés.
Gradiente de pH inmóvil	Zona de pH que es generada por diferentes anfolitos que han sido inmovilizados en un gel de poliacrilamida.
Hemicigoto	Se refiere a la condición de los hombres con respeto a los marcadores en el cromosoma X, ya que al tener solo una copia quedan como homocigotos.
Heredabilidad	Porcentaje de la variabilidad fenotípica que puede ser explicado por la variabilidad genética de los individuos.
Heredabilidad desconocida	Porcentaje de la variabilidad fenotípica que se ha atribuido a la variabilidad genética de las poblaciones, pero de lo cuál no se ha obtenido suficiente información.
Heterocigotos	Es cuando los dos alelos o formas del gen (o marcador), son diferentes, ejemplo: Aa
Heterogeneidad fenotípica	Cuando diferentes alteraciones genéticas en el mismo gen o marcador, pueden contribuir al desarrollo de diferentes variaciones de la enfermedad.
Heterogeneidad genotípica	Cuando un mismo fenotipo o enfermedad es causado por varios alelos o mutaciones (por si solos).
Hidrofobicidad	Propiedad de una molécula al no ser miscible en agua
Homocigotos	Es cuando los dos alelos o formas del gen (o marcador), son iguales, ejemplo: AA
Identidad en estado (IBS)	Es una medida de la proporción de alelos compartidos entre cada par de individuos. IBS = 1 indica que el par de muestras comparten el 100% de los alelos y por tanto se consideran duplicadas.
Identidad por descendencia (IBD)	Es una medida de la proporción de alelos compartidos entre cada par de individuos, que son heredados del mismo ancestro. Se calcula a partir de los datos IBS. Un IBD = 1 indica que el par de muestras son duplicados o que corresponden a gemelos monocigotos, un IBD = 0.5 indica que el 50% de los alelos entre el par de muestras provienen del mismo ancestro lo que indica que son parientes en primer grado, un IBD = 0.25 indica que son muestras emparentadas en segundo grado y un IBD = 0.125 indica que son muestras emparentadas en tercer grado.
Inestabilidad cromosómica	Es un tipo de inestabilidad genética que caracteriza al inicio y progresión tumoral. Se refiere a todas las alteraciones de número o de la estructura de los cromosomas.

Concepto	Definición
Inestabilidad de microsatélites (MSI)	También es un tipo de inestabilidad genética implicada en cáncer. Se caracteriza por un elevado número de mutaciones o alteraciones genéticas ganadas por errores durante la replicación en zonas altamente repetitiva del genoma, y que no son reparadas por fallas en el sistema de reparación por mal apareamiento.
Inferencia de la ancestría local (LAI)	Consiste en métodos para deducir de manera probabilistica, de que ancestro se heredo cada region del genoma a estudio, esto en base a poblaciones de referencia que se sabe han influido en la generación de la población mezclada a estudio.
Inhibición selectiva	Se refiere a la habilidad de suprimir una respuesta específica. En farmacología, los inhibiores selectivos de COX-2 son antiinflamatorios que específicamente actúan inhibiendo solo la vía COX-2, sin alterar otras fuciones, por lo cual, tiene menos efectos secundarios.
Isoelectroenfoque	Es una técnica para separar moléculas cargadas. Las moléculas se desplazan en un gradiente de pH bajo la influencia de una diferencia de potencial.
Loci	Es el plural de locus.
Locus	Se refiere a un punto en el genoma, un gen o un marcador o una región, que se esta estudiando.
Marcadores biparentales	Marcadores genéticos ubicados en los autosomas y en el cromosoma X.
Marcadores uniparentales	Marcadores genéticos ubicados en el cromosoma Y y en el mtDNA.
Oncogenes (protooncogenes)	Un protooncogén es un gen que en condiciones normales tiene como función promover el crecimiento, la proliferación y la supervivencia celular. Un oncogén es la versión mutada de los protooncogenes, por lo que estas funciones están aumentadas y descontroladas, favoreciendo así la transformación de una célula normal en una maligna y el desarrollo de cáncer.
Patrones haplotípicos	Patrones de variación a nivel genómico determinado por los bloques haplotípicos inferidos a partir de los genotipos faseados.
Penetrancia	Término estadístico que hace referencia a la frecuencia con que un fenotipo se manifiesta cuando existe el alelo de riesgo. Si 9 de 10 individuos portadores del alelo/variante/mutación de riesgo expresan el fenotipo (o la enfermedad), entonces se dice que la penetrancia es del 90%. La penetrancia puede ser entonces alta o baja: • Mutaciones germinales en genes de alta penetrancia están asociadas a enfermedades monogénicas (patrón de herencia mendeliana), por ejemplo: mutaciones en el gen BRCA1 cuya penetrancia a los 70 años es del ~ 80%. • Variantes genéticas en genes de baja penetrancia, como en casos familiares o de tipo esporádico. En estos casos, el riesgo de la variante es bajo, pero sumado a otros factores (ej: estilo de vida o dieta), van aumentando el riesgo de desarrollar la enfermedad.
Pérdida de heterocigocidad	Se refiere a la pérdida del alelo normal en un locus heterocigoto que contiene un alelo normal y uno deletereo. Al darse este fenómeno de perdida del alelo normal, el individuo pasa a ser homocigoto para el alelo mutado.

Concepto	Definición
Población efectiva	Hace referencia a la proporción de mujeres y proporción de hombres que dieron origen a la siguiente generación. En una población inicial con igual número de hombres que de mujeres, la razón de cromosoma X respecto a los autosomas es de 3/4. Sin embargo, factores como variaciones en el éxito reproductivo, las tasas de migración y diferencias en el número de hombres y mujeres, pueden cambiar la contribución de cromosomas X respecto de la de autosomas. Otros factores, como los sociales, pueden igualmente generar diferencias en la contribución parental.
Poligénico	Cuando múltiples genes o variantes genéticas, no correlacionadas, influyen en el desarrollo de un fenotipo.
Polimorfismos de un solo nucleótido (SNPs)	Cambios de secuencia de un solo nucleótido en el DNA. Son el tipo más común de polimorfismos en el genoma y en su mayoría son bialélicos. Si frecuencia en la población general es ≥ 1%
Punto isoeléctrico	Se refiere al pH al cual una molécula tiene igual número de cargas positivas que de cargas negativas, es decir, tiene carga neta cero.
Recombinación	Evento importante en genética que contribuye a la generación de nuevas combinaciones en la secuencia de DNA y, por tanto, aumenta la diversidad genética y fenotípica de una generación a la otra. Intercambio entre secuencias de DNA de miembros de un par cromosómico (homólogos) que ocurre usualmente en meiosis.
Selección natural	Proceso que permite la adaptación de las especies a su entorno. Individuos con determinadas características genéticas que tienen una mayor tasa de sobrevida o reproductiva, pasan esas características genéticas a su descendencia y por lo tanto, permanecen. Por otro lado, aquellas características genéticas con efectos deletéreos en la sobrevida o reproducción se van a perder en la siguiente generación.
Sensibilidad	Es la probabilidad de que un sujeto enfermo tenga un resultado positivo en una prueba, por ejemplo de diagnóstico. Es decir, es el porcentaje de verdaderos positivos (enfermos).
Sesgos de género	Se refiere a la diferencia en la contribución de hombres y de mujeres que participaron en la generación de una población híbrida entre dos o mas poblaciones con diferencias raciales.
Síndromes hereditarios	Conjunto de enfermedades que se explican por mutaciones de alta penetrancia heredadas por patrones mendelianos (recesivo, dominante o ligado al cromosoma X). Son usualmente monogénicas.
Sistema de desolvatación	Método para evaporar el solvente de los iones formados en la interfase que conecta el sistema HPLC con el MS. Esto contribuye a generar un espacio con alto vacío, necesario para la conducción de los iones al analizador de masas.
Subestructura poblacional	Se refiere a la ausencia de apareamiento al azar en una población. En general, la estructura poblacional habla de la distribución de individuos (homogénea o no) y del flujo génico entre los individuos de diferentes áreas, lo cual es muy importante en términos evolutivos. En estudios de asociación genéticos de casos y controles, es un concepto que se debe medir para evaluar si los dos grupos son comparables o provienen de diferentes subpoblaciones.

Concepto	Definición
tagSNP	Un SNP que por si solo puede definir la diversidad de un haplotipo, por estar ligado a otros marcadores dentro del bloque haplotípico.
Tamaño del efecto (de una variante genética)	Medida de que tanto una variable genética contribuye en el desarrollo de un fenotipo
Tasas de heterocigocidad	Medida de la diversidad en un locus polimórfico. Habla de la proporción de genotipos heterocigotos en un individuo.
Transformación logit	Método matemático para la transformación de datos (tipo proporciones o porcentajes), que no muestran una distribución normal en análisis binomiales o multinomiales. El efecto de la transformación logit es expander las diferencias entre los porcentajes de los extremos (los cercanos a 0 y los cercanos a 1), con el fin de obtener una distribución de los datos mas normal.
Transición epitelial-mesenquimal	Proceso que ocurre cuando las células epiteliales diferenciadas pierden sus características y se desdiferencian, adquiriendo propiedades mesenquimales, como: pérdida de adhesión celular, aumento en movilidad e invasividad, resistencia a apoptosis, y cambios morfológicos.
Valor predictivo positivo	La probabilidad de que una prueba con resultado positivo, corresponda a un verdadero positivo.
Variante no sinónima	Variante genética que genera un cambio de aminoácido.
Variantes funcionales	Variante genética que tiene un efecto biológico.
Variantes polimórficas	Variantes genéticas comunes en la población general (≥ 1%).

Introducción

En el mundo, el cáncer colorrectal (CCR) es el tercero más incidente en hombres y segundo en mujeres, y es una de las principales causas de mortalidad en ambos sexos (1). En Colombia es el cuarto cáncer más común en hombres y el tercero en mujeres y, ocupa el cuarto lugar entre las causas de muerte por cáncer en el país, por lo que es considerado un problema de salud pública en Colombia (1).

Durante el periodo de 1985 a 2006, la mortalidad por CCR mostró un incremento anual promedio, estadísticamente significativo, del 2,2% y 1,9% entre hombres y mujeres, respectivamente (2). Este aumento de la mortalidad fue variable entre las diferentes regiones del país (2), lo que puede estar en relación con factores ambientales y dietarios propios de las regiones, así como con factores genéticos, incluyendo la **ancestría genética**, que por nuestra historia sabemos está influenciada por componentes europeos, amerindios y africanos, y que estas proporciones son variables a lo largo del territorio colombiano (3-12).

Teniendo claro que tanto factores ambientales como de susceptibilidad genética, y su interacción, influyen de manera importante en el desarrollo de la enfermedad (13), el CCR se considera una enfermedad compleja. Desde el punto de vista genético se clasifica como hereditaria y esporádica (14), con una distribución del efecto alélico en forma de L (15). Como se muestra en la **Figura 1**, cuando la enfermedad se explica por un patrón de herencia mendeliana se caracteriza por estar asociada a variantes genéticas muy raras en la población general, o mutaciones, que tienen una **frecuencia del alelo menor, MAF** < 1% con un **tamaño del efecto** muy alto, y que al ocurrir en el material genético de las células de línea germinal, son transmitidas de generación en generación (16). Este es el caso de los **síndromes hereditarios**, causados por mutaciones germinales en genes de alta **penetrancia** como *APC* (17), *MYH* (18, 19), familia de genes *MMR* (20), *SMAD4* (21, 22), *BMPR1* (23), *STK11* (24), *TP53* (25) y

PTEN (26), que explican solo hasta un 5% de todos los casos de CCR (ver **Tabla 1**) . Por otro lado, el CCR esporádico se caracteriza por una importante **heterogeneidad genética**, al estar asociado a un gran número de **variantes polimórficas** que son frecuentes en la población general (MAF > 1%) pero que tienen un tamaño del efecto bajo (16). Es decir, consiste en una enfermedad **poligénica**, en la que no una sola variante en un gen, sino la sumatoria de muchas variantes en varios genes de baja penetrancia, y su interacción entre sí y con otros factores relacionados con el estilo de vida y la dieta, modifican el riesgo de desarrollar la enfermedad (14).

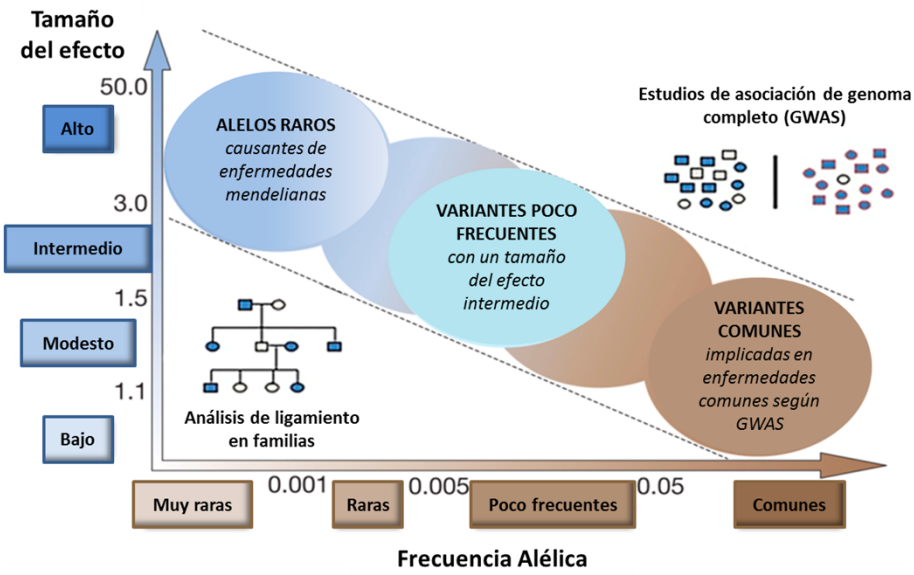


Figura 1 Arquitectura alélica en enfermedades complejas y los métodos de abordaje para el estudio de marcadores genéticos asociados al riesgo de enfermedad. Figura modificada de *Manolio T, et al (16).*

El adenocarcinoma comprende más del 95% de todos los CCR y éstos, predominantemente, se considera son derivados de lesiones tumorales precursoras de tipo epitelial, conocidos como pólipos adenomatosos (PA) o adenomas (27), que consisten en lesiones displásicas e hiperproliferativas. La malignización del adenoma hacia adenocarcinoma, está bien documentado como un proceso de múltiples eventos que involucra alteraciones en **genes supresores de tumor** y **oncogenes**, con la consecuente alteración en el control del ciclo celular (13). La historia natural de este proceso, conocido como vía de progresión adenoma-adenocarcinoma, se da en un periodo de 10 a 15 años, principalmente por dos vías, la supresora y la mutadora (13, 28,

29), que enmarcan la **heterogeneidad fenotípica** característica de la enfermedad con diferentes implicaciones clínicas[1].

Tabla 1 Genes de alta penetrancia asociados a síndromes hereditarios de cáncer colorrectal, estratificado según las vías de carcinogénesis implicadas

GEN	CLASE	FUNCIÓN PROTEICA	ENFERMEDAD ASOCIADA
Vía Supresora o de Inestabilidad Cromosómica			
APC (Adenomatous Polyposis Coli)	Supresor de tumor	Regula la degradación de β-catenina. Regula la segregación de cromosomas en mitosis-anafase.	1. Poliposis Adenomatosa Familiar (FAP) 2. Síndrome de Gardner 3. Cáncer esporádico gastrointestinal
SMAD4 (Mothers Against Decanpentaplejic Pathway Homolgs 4)	Supresor de tumor	Regula la vía de señalización de TGF-β.	1. Síndrome de Poliposis Juvenil (JPS) 2. Síndrome de Poliposis Juvenil combinado con Telangiectasia Hemorrágica Hereditaria (JP-HHT)
BMPR1 (Bone Morphogenetic Protein Receptor 1)	Supresor de tumor	Proteína transmembranal tipo I (receptor tirosina/cinasa) que participa en la señalización mediada por las BMPs.	1. Poliposis Juvenil 2. Síndrome hereditario de poliposis mixta
STK11 (Serine/Threonine Kinase 11)	Supresor de tumor	Regula la polaridad celular.	1. Síndrome de Peutz-Jeghers 2. Pólipos hamartomatosos del colon
PTEN (Phosphatase and Tensin Homolog)	Supresor de tumor	Fosfatasa de PIP_3. Regula negativamente la vía PI3K/Akt de supervivencia celular.	1. Síndrome de Cowden 2. Cáncer hereditario gastrointestinal y mamario 3. Cáncer esporádico de tiroides y próstata 4. Glioblastoma
TP53 (Tumour Protein 53)	Supresor de tumor	Factor de transcripción que controla el ciclo celular al permitir su arresto para favorecer la reparación del DNA; es proapoptótico.	1. Síndrome de Li-Fraumeni 2. Varios cánceres esporádicos.
Vía Mutadora o de Inestabilidad de Microsatélites			
MYH (MutY Homolog)	Cuidador o Mutador	Reparar errores del DNA por daño oxidativo, mediante el mecanismo de escisión de bases (BER).	1. Poliposis Asociada a MUTYH
Familia de genes *MMR* (Mismatch Repair Genes)	Cuidador o Mutador	Repara malos apareamientos en el DNA.	1. Síndrome de Lynch I (Cáncer Colorrectal No Polipósico Hereditario, HNPCC) 2. Síndrome de Lynch II (incluye el síndrome de Muir-Torre)

[1] Para más información con respecto a las vías de carcinogénesis del CCR y los eventos moleculares involucrados, remitirse al artículo de revisión preparado y publicado como producto de esta tesis doctoral: *Sanabria MC, et al. Vías de carcinogénesis colorrectal y sus implicaciones clínicas. Revista Colombiana de Cancerología. 2012:16(3):170-181*

La vía supresora es la más común, pues representa el 85% de los CCR de tipo esporádicos. Esta vía también explica los casos de Poliposis Adenomatosa Familiar, FAP; afecta más frecuentemente el colon distal y se caracteriza por generar **inestabilidad cromosómica**, **aneuploidía** y **pérdida de heterocigocidad** (ver **Tabla 1** y **Tabla 2**) (30-32). La vía mutadora es característica del 15% de los CCR esporádicos y también explica los casos de Síndrome de Lynch I o CCR hereditario no polipósico, HNPCC (ver **Tabla 1**). Esta vía es más común en colon proximal y se caracteriza por presentar un **fenotipo mutador** en regiones microsatélite, conocido como **inestabilidad de microsatélites, MSI**, acompañado de estabilidad cromosómica, es decir, los tumores son **diploides** (ver **Tabla 2**) (28).

Tabla 2 Diferencias embriológicas, morfológicas y moleculares del intestino grueso

CARACTERÍSTICAS		COLON PRÓXIMAL	COLON DISTAL
Tejido Normal			
Embriológicas		Intestino medio	Intestino caudal
Morfológicas	*Clasificación*	Derecho (ciego al ángulo esplénico)	Izquierdo (colon descendente al recto)
	Irrigación	Arteria mesentérica superior	Arteria mesentérica inferior
	Red capilar	Varias capas	Una sola capa
Moleculares	*Expresión de Bak*	Baja	Alta
	Apoptosis	Baja	Alta
Tejido Tumoral			
Morfológicas	*No. de células caliciformes*	Menor	Mayor
	Producción de moco	Baja	Alta
Moleculares	*Cariotipo*	Diploidía	Aneuploidía
	Inestabilidad Genética	Inestabilidad de Microsatélites	Inestabilidad Cromosómica
	Metilación del DNA	Alta	Baja
Vías de Carcinogénesis			
Clásicas		Vía Mutadora	Vía Supresora

Hasta el 90% de los pacientes con CCR en etapas tempranas se curan con cirugía (14), pero desafortunadamente es diagnosticada en su mayoría de veces en estadios avanzados cuando el pronóstico es pobre. Actualmente, no existe una prueba para el diagnóstico temprano de CCR con alta **sensibilidad**, **especificidad** y **valor predictivo positivo**; el diagnóstico del CCR aún se basa solo en la clasificación descriptiva y

sistemas de estadificación de acuerdo a la evaluación morfológica e histológica del tumor (33). Teniendo en cuenta lo anterior, sería de gran aporte contribuir a la identificación de biomarcadores que sirvan no solo en el diagnóstico de la enfermedad en estadios tempranos, disminuyendo así las tasas de mortalidad por CCR, sino también para estratificar las poblaciones en mayor riesgo de desarrollar la enfermedad, en los cuales sea posible implementar medidas de tamizaje y en general de prevención primaria. La búsqueda continua de biomarcadores de mayor susceptibilidad para el desarrollo de CCR y de detección temprana, se ha promovido con los avances en diferentes áreas de las "Ómicas" con el fin de explicar mejor las diferencias relacionadas con la heterogeneidad genética y fenotípica de la enfermedad. Lo anterior incluye la implementación de análisis de polimorfismos genéticos (34-56), expresión génica (57-61), expresión proteica (62-65) y de regulación epigenética (66, 67), todas estas herramientas útiles para el establecimiento de patrones moleculares que sugieran posibles estrategias de control de la enfermedad en la práctica clínica.

En este contexto, en este trabajo de tesis se implementaron herramientas a gran escala como microarreglos de DNA para la búsqueda de marcadores de susceptibilidad en el desarrollo de CCR en poblaciones colombianas, tipo **polimorfismos de un solo nucleótido, SNPs**, que puedan ser usados en la estratificación a nivel poblacional de individuos en mayor riesgo de la enfermedad. Adicional a lo anterior, se implementaron métodos en proteómica con el fin de contribuir a la caracterización de los perfiles de expresión proteica de casos y controles incluidos en el estudio, como una aproximación para la búsqueda de biomarcadores en sangre que sean de utilidad en el diagnóstico temprano y pronóstico de la enfermedad. El principal impacto de éste trabajo es el de avanzar en el conocimiento de las bases biológicas y moleculares del desarrollo tumoral de colon y recto en la población colombiana, con el fin de proponer marcadores que puedan ser posteriormente validados y usados en el control de la enfermedad a través de acciones en salud pública.

Los resultados de este trabajo de tesis están enmarcados en tres objetivos generales y cada uno se desarrolla en un capítulo, de la siguiente manera:

CAPÍTULO	OBJETIVOS GENERALES	PROPÓSITOS	OBJETIVOS ESPECÍFICOS
1	**1. Evaluar el papel de la estructura genética de las poblaciones colombianas en el riesgo de tumores colorrectales**	*Medir objetivamente las diferencias en las proporciones de ancestría en las poblaciones incluidas y evaluar posibles sesgos de género que hicieron parte del proceso de mestizaje.*	1. Estimar las proporciones de ancestría globales a nivel de genoma completo, autosomas y cromosoma X, en colombianos incluidos en el estudio. 2. Evaluar diferencias en las estimaciones en cromosoma X versus autosomas para la población colombiana, y discriminado por regiones y por ciudades 3. Calcular la proporción de mujeres de ancestría europea, amerindia y africana, que contribuyeron al proceso de mestizaje de las poblaciones incluidas en el estudio
		Evaluar la influencia de la composición genética ancestral heterogénea de los colombianos en la modificación del riesgo de desarrollar PA y CCR	1. Describir las características de la totalidad de la muestra de casos y controles colombianos incluidos en el estudio 2. Evaluar diferencias en las proporciones de ancestría globales a nivel de autosomas entre casos y controles colombianos 3. Evaluar el efecto de factores no genéticos en el riesgo de PA y CCR en colombianos con diferentes proporciones de ancestría
2	**2. Identificar variantes genéticas comunes asociadas al riesgo de tumores colorrectales en los colombianos**	*Replicar algunas variantes genéticas previamente asociadas con CCR, en la muestra de casos y controles colombianos del estudio.*	1. Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, de 14 SNPs previamente publicados, mediante análisis básicos de asociación por SNP (X^2) y regresiones logísticas ajustadas.
		Descubrir nuevas variantes genéticas comunes, asociadas al riesgo de PA y CCR en colombianos, tomando ventaja del alto grado de mezcla de nuestra población.	1. Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, usando una aproximación de genes candidatos (CG) 2. Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, usando una aproximación de genoma completo (GWAS)
3	**3. Identificar perfiles proteómicos en colombianos con cáncer colorrectal**	*Explorar diferencias en los perfiles proteómicos en plasma entre pacientes con CCR y controles.*	1. Evaluar diferencias en la expresión de proteínas en plasma entre pacientes con CCR y controles del estudio. 2. Proponer proteínas que actúen como posibles biomarcadores de diagnóstico o pronóstico para CCR

Para cada objetivo general se describen uno o dos propósitos, dentro de los cuales se establecieron los objetivos específicos respectivos. Cada propósito cuenta con una breve descripción metodológica, que puede ampliarse revisando los *Anexos* respectivos. Los resultados responden directamente a cada objetivo específico y estos se discuten al final de cada capítulo.

1. Estructura genética de poblaciones colombianas y su papel en el riesgo de pólipos adenomatosos y cáncer colorrectal

Antecedentes de la colonización en Colombia:

Los datos genéticos obtenidos del análisis de **DNA antiguo (aDNA)**, han favorecido el estudio de la evolución de la especie humana y de sus migraciones alrededor de todo el mundo, permitiendo entender mejor la diversidad genética de las poblaciones modernas (68, 69). Estos avances han resultado muy importantes en el diseño adecuado de estudios poblacionales, dirigidos a la identificación de factores genéticos de riesgo en el desarrollo de enfermedades complejas (70).

Actualmente, es aceptado que América fue el último de los continentes en ser habitado por humanos, gracias a una migración mayor desde Siberia a través del estrecho de Bering, que según evidencia genética y arqueológica tomó lugar desde hace 15.000 a 12.000 años (ver **Figura 2**) (69). Desde allí, estos grupos se distribuyeron a lo largo de Norte América avanzando posteriormente a territorio Sur Americano, muy posiblemente, a través de Panamá y del departamento del Chocó en Colombia (71). De acuerdo con los estudios basados en los refugios de piedra, del antropólogo Tom Dillehay, los primeros humanos que habitaron el país se concentraron a lo largo de la Costa Caribe y de las laderas de los Andes (72). Este movimiento y asentamiento de los nativos americanos o amerindios en Colombia se dió entre 9.000 a 1.500 años a.C. y consistieron en tres grupos principales: los Chibchas que dieron origen a los *Muiscas* ubicados en las áreas de Cundinamarca y Boyacá, y a los *Taironas* que se ubicaron en la Sierra Nevada de Santa Marta, entre otros; los Caribe que incluye a los *Chocoes* de la Costa Pacífica y a los *Quimbayas* del occidente de la Cordillera Central de los Andes, entre otros; y los Arawak

que originaron a los *Guajiros* y *Guahíbos* ubicados en la Guajira y Llanos Orientales, respectivamente, entre otros (73).

Poco después del descubrimiento de América en el año 1492, entraron los primeros europeos provenientes de la península Ibérica a Colombia, a través de Cartagena, desde donde se extendieron por el resto de Sur América a inicios del siglo XVI (ver **Figura 2**) (74). Es así como las primeras ciudades colombianas aún vigentes, fundadas por españoles, fueron Santa Marta en 1525 y Cartagena en 1533, convirtiéndose éste último en el principal puerto del sur del continente (74). Una expedición al interior del país a lo largo del río Magdalena favoreció el asentamiento de españoles en lo que ahora conocemos como Bogotá, antigua nación Muisca, y hacia el final del siglo XVI ya se habían fundado la mayoría de las principales ciudades de Colombia (74). Casi simultáneamente con la conquista española hasta el siglo XIX, se inició la trata de esclavos provenientes del occidente de África al continente Sur Americano (ver **Figura 2**) (70); los hombres negros fueron obligados a hacer trabajos pesados en las minas de oro de los Andes Occidentales, y de la Costa Pacífica, y en las haciendas de la llanura costera del Caribe, mientras que las mujeres negras trabajaron principalmente en servicio doméstico para familias españolas en zonas urbanas del país (75).

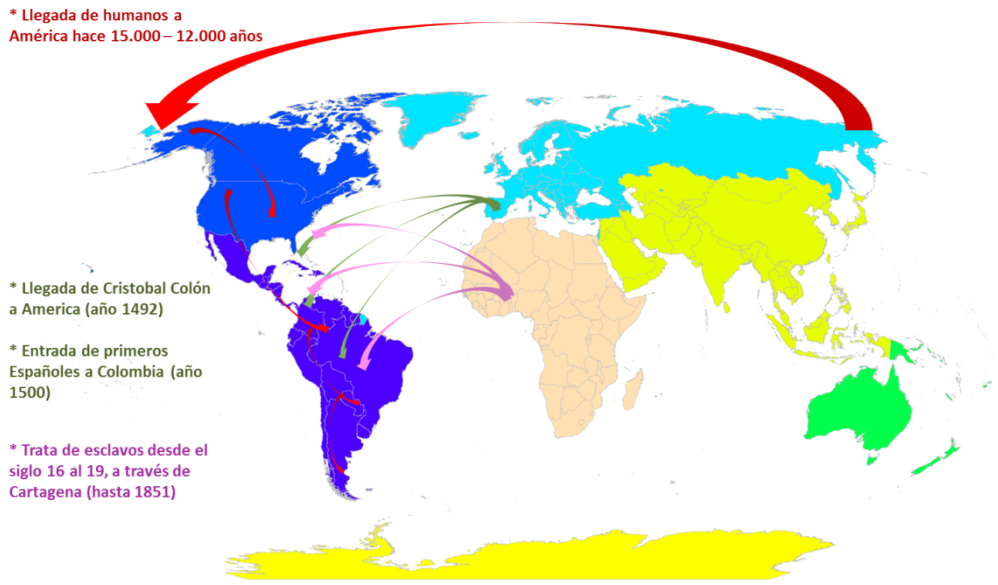


Figura 2 Antecedentes de la diversidad genética en Sur América, caracterizadas por la mezcla de ancestros amerindios, europeos y africanos. Figura original; mapa generado en el programa *R statistics* (76).

La alta mortalidad indígena a inicios de la colonización generó la necesidad de repoblar el territorio, dando origen a diferentes grupos étnicos a lo largo del país con grados de mezcla variables entre europeos, amerindios y africanos establecidos en el territorio colombiano desde el siglo XVI. Dentro de éstos se encuentran: i) los *criollos*, que corresponden a descendientes de europeos por padre y madre, nacidos y criados en el continente americano; ii) los *mestizos*, producto de la mezcla entre europeos y amerindios, siendo estos el grupo demográfico más numeroso desde finales de la colonización hasta tiempos actuales; y iii) las poblaciones afrodescendientes, que incluye *palenqueros o cimarrones* que escaparon de la esclavitud y se ubicaron en regiones aisladas de las Costas Caribe y Pacífica formando colonias con muy poco grado de mezcla, los *afrocolombianos* originados a partir de la mezcla de africanos con poblaciones europeas, amerindias y mestizas, que se ubicaron principalmente en la costa Caribe y Pacífica, y los *raizales* de origen angloafricanos que se ubicaron principalmente en las islas de San Andrés y Providencia (10, 77).

En la actualidad, las poblaciones con mayor porcentaje de individuos auto-catalogados como amerindios, están ubicadas en el Vaupés, Guainía, Guajira, Vichada, Amazonas, Cauca y Putumayo, con porcentajes que van desde 67% a 21%; mientras que en los departamentos de la Guajira, Cauca, Nariño, Córdoba y Sucre se concentran el 65,77% de los indígenas del país (78-80). Por otro lado, las poblaciones afrodescendientes están asentadas en el Chocó, San Andrés, Bolívar, Valle del Cauca, Cauca, Nariño y Sucre, con porcentajes que van desde 82% a 16%; mientras que los departamentos en donde se concentra aproximadamente el 50% de la población negra del país son Valle del Cauca, Antioquia y Bolívar, siendo Cali la ciudad donde reside actualmente el mayor porcentaje de población afrocolombiana del país, seguido por Cartagena y Buenaventura (79-82). Otras regiones como Bogotá, Antioquia, Barranquilla y Santander, fueron los principales asentamientos de familias europeas gracias a eventos migratorios posteriores que incluyeron una mayor proporción de mujeres; por tanto, éstas poblaciones son consideradas como los más "blancos" del país.

Teniendo en cuenta lo anterior, es de esperarse que Colombia sea un país con una gran diversidad étnica y sociocultural, favorecida además por su geografía, que va desde las Costas Pacífica y Caribe, hasta las tres Cordilleras de los Andes, dividiendo el país en tres zonas: i) desde la Costa Pacífica hasta la Cordillera Occidental; ii) la Cordillera

Central que abarca desde el Cauca hasta el río Magdalena; y iii) desde la Cordillera Oriental hasta los límites con Venezuela. Vale la pena resaltar que ésta distribución geográfica, y sus diferencias climáticas características, moldearon el estilo de vida y la dieta de sus pobladores, dando lugar al establecimiento de diferencias regionales muy marcadas a lo largo del territorio nacional.

Evidencia genética de la heterogeneidad en el proceso de mestizaje en diferentes regiones del país:

De los estudios más grandes realizados en Colombia para evidenciar las variaciones en el componente ancestral genético de nuestras poblaciones, está el realizado por *Rojas W, et al* (10). Este estudio incluyó 15 poblaciones urbanas de las regiones andina y costera, 8 poblaciones de amerindios y 1 población de afrocolombianos; usando **marcadores biparentales**, 8 en **cromosoma X** y 11 en **autosomas**, estimaron las proporciones de ancestría y encontraron una importante heterogeneidad entre las regiones, tanto a nivel de autosomas como del cromosoma X (10). En términos generales, a nivel de autosomas encontraron que en poblaciones urbanas las proporciones de ancestrías europea, amerindia y africanas fueron aproximadamente 42%, 47% y 10%, con porcentajes mayores de ancestría africana en población negra del Chocó, Bolívar, Magdalena, Cauca y Valle del Cauca (68%, 44%, 28%, 24% y 22%, respectivamente) (10). El componente amerindio se observó mayor que el europeo en poblaciones de Bolívar, Peque, Norte de Santander, Casanare, Nariño, Cauca y el Huila; mientras que un porcentaje de ancestría europea mayor que la amerindia se evidenció en el Magdalena, Antioquia, Caldas, Quindío, Santander y negros del Chocó (10). Igualmente, en el estudio de *Rojas W, et al.,* encontraron que la ancestría amerindia calculada en cromosoma X fue mayor a la calculada en autosomas en todas las poblaciones (10). Más aún, a nivel del cromosoma X en todos los casos, a excepción de Antioquia, la ancestría amerindia fue mayor al componente europeo (3, 4, 6, 10, 11).

En la **Tabla 3**, se resumen los resultados de múltiples estudios en varias poblaciones colombianas en los cuales se estimaron las proporciones de ancestrías europea, amerindia y africana, usando marcadores en autosomas y en cromosoma X.

Tabla 3 Estimaciones de la ancestría en autosomas y cromosoma X en diferentes poblaciones colombianas

POBLACIONES COLOMBIANAS		PROGRAMAS	ESTIMACIONES DE ANCESTRÍA (%)								REFERENCIAS
			AUTOSOMAS				CROMOSOMA X				
Muestras	N		Marcadores	Europea	Amerindia	Africana	Marcadores	Europea	Amerindia	Africana	
Antioquia	80	ADMIX	8 polimorfismos	79	16	6	5 polimorfismos	69	25	6	Bedoya G, et al. 2006
Antioquia	30	STRUCTURE	1536 SNPs	71	19	10	-	-	-	-	Price AL, et al. 2007
Antioquia (Medellín)	20	STRUCTURE	678 polimorfismos	66	25	9	29 polimorfismos	53	29	19	Wang S, et al. 2008
Cundinamarca	19			46	51	3		26	58	16	
Nariño (Pasto)	19			39	57	4		33	53	14	
Peque	20			37	58	5		31	56	13	
Bolivar	80	STRUCTURE	11 Indels-SNPs	23	33	44	8 STRs	20	71	5	Rojas W, et al. 2010
Magdalena	26			50	22	28		38	75	0	
Antioquia	80			64	26	11		45	44	11	
Peque	163			32	62	6		-	-	-	
Caldas	203			59	36	4.3		56	63	0	
Quindío	58			57	38	4		21	97	0	
Norte de Santander	35			42	53	5		32	79	0	
Santander	82			56	42	1		67	78	0	
Cundinamarca	24			45	52	3		5	88	8	
Casanare	20			25	75	1		65	82	0	
Nariño	201			32	65	3		0	96	7	
Cauca	61			20	57	24		0	76	34	
Huila	24			40	61	0		0	70	30	
Valle del Cauca	124			39	39	22		31	79	0	
Chocó	170			47	45	9		5	87	7	
Afro Chocó	72			21	11	68		31	59	11	
Nativos Americanos (Awa)	22	ADMIXTURE	325 SNPs	17	80	2	-	-	-	-	Galanter JM, et al. 2012
Nativos Americanos (Coyaima)	19			9	86	2					
Nativos Americanos (Pasto)	36			16	83	2					
Antioquia	19			52	39	6					
Chocó	35			10	13	76					
Mulaló	28			25	18	54					
Cauca	306	ADMIXMAP	34 Indels-SNPs	48	41	11	9 polimorfismos	32	57	11	Córdoba L, et al. 2012
Región Caribe (Cartagena)	194	STRUCTURE	52 SNPs	-	-	-	32 Indels	31	34	36	Ibarra A, et al. 2014
Costa Pacífica (Chocó)	104			23	23	53		28	28	44	
Region Andina Occidental (Antioquia)	145			44	35	21		38	34	29	
Region Andina Oriental (Boyacá, Santander, Norte de Santander, Huila)	500			42	39	20		35	38	27	
Region de la Orinoquía (Arauca, Casanare, Meta)	231			41	38	22		33	38	29	
Region Andina Sur-Occidental (Pasto)	84			30	51	19		24	51	25	
Nativos Americanos (Antioquia)	38	STRUCTURE	46 Indels	3	95	2	-	-	-	-	Xavier C, et al. 2015
Nativos Americanos (Cauca)	55			17	78	5					
Antioquia	60	ADMIXTURE	Genoma Completo	75	18	7	-	-	-	-	Rishishwar L, et al. 2015
Nativos Americanos (Guainía)	35	STRUCTURE	46 Indels	4	94	2	-	-	-	-	Ossa H, et al. 2015
Nativos Americanos (Motilón-Barí)	57			4	95	1					
Nativos Americanos (Pijaos)	29			32	61	7					

Como es de esperarse dada nuestra historia desde la colonización y de acuerdo a los resultados de *Rojas W, et al.*, existen diferencias en las estimaciones de ancestría entre las regiones colombianas (10) (ver **Tabla 3**). Sin embargo, también se observan algunas variaciones en las proporciones ancestrales para una misma región entre los estudios, lo cual podría deberse tanto a las diferencias en el número y capacidad discriminatoria de los marcadores analizados, como al programa bioinformático implementado.

Evidencia genética de sesgos de género en el proceso de mestizaje y su influencia en el moldeamiento de la estructura genética variable en las regiones colombianas:

En Colombia, los primeros estudios genéticos para evaluar si durante el proceso de mestizaje existieron **sesgos de género** o diferencias en la contribución de hombres y mujeres de cada ancestría en el proceso de mezcla, se realizaron usando **marcadores uniparentales** de la región no recombinante del **cromosoma Y** y del **DNA mitocondrial (mtDNA)** en el cual no existe **recombinación**; lo anterior, con el fin de estimar el componente ancestral paterno y materno, respectivamente, de quienes dieron origen a la población actual en Antioquia (70). Con respecto al linaje paterno, se encontró que el 94% de los cromosoma Y eran de origen europeos, el 5% eran africanos y el 1% eran amerindios; mientras que en relación al linaje materno, observaron que el 90% de los mtDNA eran de origen amerindios, el 8% eran africanos y el 2% eran europeos (70).

De acuerdo a estos primeros reportes, la población antioqueña es el resultado de una asimetría en el proceso de mestizaje que incluyó, principalmente, hombres europeos con mujeres amerindias, lo que está en relación con lo encontrado en otras poblaciones de Colombia como Caldas, Quindío, los Santanderes, Cundinamarca, Casanare, Nariño, Cauca, Huila, Valle del Cauca y población mestiza del Chocó, todos con mtDNA amerindios en más del 84%; mientras que otras poblaciones como Bolívar y población negra del Chocó, tuvieron porcentajes de mtDNA amerindios del 56% y 47%, con su contraparte africana del 21% y 53%, respectivamente (83), indicando que en éstas poblaciones también hubo una representación importante de mujeres de origen africano en el proceso de mestizaje.

Igualmente, con respecto a la proporción de cromosomas Y de predominio europeo en los antioqueños, se observaron resultados similares en otro estudio para Antioquia y otras poblaciones como Caldas, Santander, Casanare y Huila, con porcentajes mayores al 85% (83). A su vez que se evidenció una participación representativa de cromosomas Y de origen africano en otras poblaciones como Valle del Cauca (33%), Bolívar (26%) y población negra del Chocó (24%); mientras que se observó una representación sustancial del linaje paterno amerindio en el Magdalena (33%), Peque (31%), Quindío (29%), negros del Chocó (24%) y Nariño (21%) (83).

Si bien estos primeros reportes en marcadores uniparentales han evidenciado, en parte, las diferencias en la contribución de hombres y mujeres de cada ancestría que participaron en el proceso de mestizaje en Colombia, es ampliamente conocido que mediante el análisis de marcadores biparentales se obtienen mejores estimados de esta asimetría de género y su papel en el modelamiento de las poblaciones actuales (3, 4, 6, 9-11, 84, 85). Esta ventaja se debe a que, a diferencia de los marcadores uniparentales, en los autosomas se dan eventos de recombinación homóloga y representan tanto la herencia materna como paterna, y también a que el patrón de herencia del cromosoma X es diferente en los dos sexos; las mujeres heredan éste cromosoma de sus dos progenitores, mientras que los hombres solo pueden heredarlo de su madre (86, 87).

Teniendo en cuenta éstas características y asumiendo una población con igual número de hombres que de mujeres, la media de la mezcla en autosomas es el promedio de las contribuciones femeninas y masculinas en la población (ver Ecuación (a), $H^{A}_{1,g,\delta}$); mientras que, en la media de la mezcla en los cromosomas X de una población, 2/3 es aportado por mujeres y 1/3 por hombres (ver Ecuación (b), $H^{X}_{1,g,\delta}$) (87).

(a) $$\mathbb{E}[H^{A}_{1,g,\delta}] = \frac{1}{2}s_1^f + \frac{1}{2}s_1^m$$

(b) $$\mathbb{E}[H^{X}_{1,g,\delta}] = \frac{2}{3}s_1^f + \frac{1}{3}s_1^m$$

Ecuación (a) y (b) Contribución de hombres ($\%_{relativo,hom} = {}^{m}$) y mujeres ($\%_{relativo,muj} = {}^{f}$) de cada ancestría (*S*, población ancestral; S_1 = europea, S_2 = amerindia, S_3 = africana) en la población mezclada a estudio (*H*). Estas contribuciones pueden ser calculadas a partir de las medias observadas o estimadas de cada ancestría (S_1, S_2 o S_3) en autosomas (H^A) y en cromosoma X (H^X) de dicha población. Lo anterior, asumiendo un evento de mestizaje en un momento en el tiempo. Tomado de *Goldberg A et al* (87).

Evidencia sobre el papel de la ancestría en el riesgo de cáncer en poblaciones con alto grado de mestizaje:

Se han reportado varios ejemplos de **disparidad** en el riesgo de cáncer que han sido explicados, en parte, por el efecto de la ancestría genética de las poblaciones. Por ejemplo, se observó que en mujeres latinas de Estados Unidos de América, USA, aumentaba el riesgo de cáncer de mama por cada aumento del 25% en la ancestría europea, aún después de corregir por factores de riesgo conocidos y lugar de nacimiento [OR 1.39; 95% CI 1.06-2.11; P = 0.013] (88); ésta asociación se replicó en un grupo independiente de mujeres mexicanas (89). Igualmente, se observó que un incremento en las proporciones de ancestría amerindia en el **locus** 6q25, contribuye a disminuir el riesgo de cáncer de mama en mujeres latinas [OR 0.75; 95% CI 0.65-0.85, $P = 1.1 \times 10^{-5}$, *P permutaciones* = 0.02] (90); más aún, dentro de este locus se logró identificar un SNP de origen amerindio que explica gran parte de esta asociación como factor protector [6q25:rs140068132, OR 0.60; 95% CI 0.53-0.67; $P = 9 \times 10^{-18}$) (91).

Con respecto a la mortalidad por cáncer de mama, se evidenció que en mujeres latinas de USA con una fracción ancestral amerindia global de más del 50%, se aumentaba hasta dos veces más el riesgo de morir por la enfermedad, con respecto a aquellas con una fracción ancestral amerindia menor al 50% (92). También, se ha observado que las mujeres afroamericanas muestran una baja incidencia global de cáncer de mama, pero las más altas tasas de mortalidad relacionadas con la enfermedad en comparación con las mujeres blancas de USA; lo anterior, aun corrigiendo por factores relacionados con el nivel socioeconómico y acceso al sistema de salud (93-95). Parte de la explicación puede estar en que se encontró que las mujeres afroamericanas con altos porcentajes de ancestría africana, son las que presentan tumores más agresivos y de mayor grado histológico que no expresan los receptores hormonales (ER-, PR-) (96, 97); en relación a esto, se ha reportado una alta prevalencia del subtipo basal en mujeres premenopáusicas afrodescendientes (98).

Con respecto a disparidades en otros cánceres, se ha reportado un mayor riesgo de cáncer de próstata en afroamericanos con porcentajes altos de ancestría africana local en la región 8q24 o con una mayor proporción de ancestría europea en la región 11p13 (99-

101). Si bien, poco se ha estudiado sobre el efecto de la ancestría en la respuesta a tratamientos en cáncer en general; *Yang, et al.,* evidenciaron un mayor riesgo de recaída después del tratamiento de quimioterapia en los pacientes pediátricos de USA, diagnosticados con leucemia linfoblástica aguda (LLA), con incrementos de más del 10% en las proporciones de ancestría amerindia global, y que la variación genómica dada por esta ancestría podría estar relacionada con resistencia al tratamiento, afectando las tasas de curación en estos pacientes (102).

En CCR, las tasas de incidencia en USA muestran diferencias raciales y étnicas. En un estudio de cohorte multiétnico a 10.7 años que incluyó alrededor de 160 mil participantes de origen afroamericano, japoneses americanos, latinos, nativos de Hawaii y blancos - no hispanos de USA, encontraron que hombres y mujeres japoneses americanos y mujeres afroamericanas mostraron un mayor riesgo de CCR en comparación con los blancos - no hispanos; más aún, las mujeres japonesas americanas y afroamericanas presentaron mayor riesgo de enfermedad avanzada que la población blanca de USA (103). Llama la atención que estas diferencias se encontraron en modelos ajustados por edad, historia familiar de CCR, antecedentes de pólipos, IMC, consumo de AINES y otros factores de riesgo no genéticos; por lo cual, este estudio concluye que parte de las disparidades étnicas observadas para CCR en USA pueden ser explicadas por diferencias en la distribución de factores de riesgo conocidos (103). Sin embargo, otros factores como la susceptibilidad genética pueden estar teniendo un rol importante en estas variaciones con respecto al riesgo (103).

La mayor susceptibilidad de CCR en población afroamericana ha sido muy estudiada, debido a que esta minoría es la que presenta las más altas tasas de incidencia y mortalidad por la enfermedad, en comparación con otros grupos étnicos en USA (104). Recientemente se reportó que la incidencia (por 100.000 habitantes) ajustada por edad a los 50 años fue de 44.2 en blancos de USA, mientras que en afroamericanos fue de 62.6; más aún, se determinó que la edad a la cual las tasas de incidencia empezaban a incrementar fueron de 47 y 43 años en blancos y negros de USA, respectivamente, y que estas diferencias se mantuvieron en los análisis estratificados por nivel socioeconómico (104). A partir de esta evidencia, sociedades como el *American College of Physicians* (105), el *American College of Gastroenterology* (106) y el *American Society for Gastrointestinal Endoscopy* (107), han incluido dentro de sus guías la recomendación de

iniciar el tamizaje en CCR a los 45 años en población afroamericana, mientras que la recomendación para la población general sigue siendo la de iniciar el tamizaje a los 50 años (104). La conclusión general es que el alto riesgo en la incidencia y mortalidad por CCR en afroamericanos en comparación con otros grupos étnicos, resulta de la combinación e influencia de factores de riesgo no genéticos con factores biológicos/genéticos de susceptibilidad; lo anterior, a pesar de la presencia de otros relacionados con el nivel socioeconómico, como son el acceso a la salud y la educación (104).

También se han reportado disparidades en el riesgo de CCR en poblaciones Latinas en USA y en sus países de origen (108). Un factor importante a tener en cuenta, es que en estas poblaciones existe un alto grado de mestizaje que además es variable a lo largo de todo el continente; esta heterogeneidad en la estructura genética junto con la influencia de otros factores no genéticos, contribuyen a las variaciones en el riesgo de CCR observado dentro de este grupo étnico (108). Estas diferencias en la estructura genética ancestral de los Latinos, lastimosamente, condicionan también diferencias en otros factores; por ejemplo, se ha reportado que Latinos con mayor proporción de ancestría global amerindia o africana, tienen menor nivel socioeconómico en comparación con Latinos con mayor ancestría europea en USA (108). Si bien las tasas de incidencia y mortalidad de CCR son menores en Latinos comparado con poblaciones blancas – no hispanas en USA, hay diferencias según el país de origen (108); en un estudio realizado en la Florida, las tasas de incidencia en CCR en hombres Cubanos y Puertorriqueños se observaron similares a las tasas en hombres blancos en USA, pero dos veces más altas que en hombres Mexicanos (109), lo cual fue confirmado por *Stern MC, et al.,* en California (110).

Antes de nuestro estudio, dentro del cual está anidada ésta tesis de doctorado, no existen reportes de asociación de los componentes de ancestría genética y el riesgo de CCR en poblaciones mezcladas como Colombia; los resultados obtenidos en los análisis de la primera mitad de la muestra[2] y del total de la muestra[3] fueron publicados por nuestro grupo de investigación. Los hallazgos completos se describen en la sección de resultados de este documento.

Evidencia sobre la variabilidad en el riesgo de CCR en poblaciones colombianas:

Se han reportado diferencias departamentales en la incidencia y mortalidad del CCR en hombres y mujeres de Colombia (ver **Tabla 4** y **Tabla 5**, respectivamente) (2, 111, 112). Las tasas estimadas de incidencia obtenidas para el periodo 2000 a 2006, fueron generadas solo a partir del Registro Poblacional de Cali, mientras que las tasas de incidencia para el periodo 2007 a 2011 fueron estimadas a partir de la información del Registro Poblacional de Cali, Pasto, Manizales y Bucaramanga (111, 112); por esta razón no es posible hacer una comparación directa entre los dos periodos, sin embargo, ambos se registran en la **Tabla 4** y **Tabla 5** para fines ilustrativos.

Durante el periodo de 2007 a 2011, la incidencia estimada anual de CCR en hombres fue mayor en el Quindío, Bogotá DC, Risaralda, Caldas, Valle del Cauca, Meta, Santander, Antioquia y Cundinamarca con tasas ajustadas por edad (TAE) de 18.3 a 13.5; mientras que los departamentos con menor incidencia fueron Bolívar, San Andrés y Providencia, Magdalena, Córdoba, Putumayo, Grupo Amazonas, Sucre, Nariño, Cauca, Caquetá, Chocó y la Guajira, con TAE de 7.8 a 3.0 (ver **Tabla 4**) (111).

---

[2] Resultados de las estimaciones de ancestría y su asociación con neoplasias colorrectales con la primera mitad de la muestra, publicado en: *Hernández-Suárez G, Sanabria MC, et al. Genetic ancestry is associated with colorectal adenomas and adenocarcinomas in Latino populations. European Journal of Human Genetics. 2014;22:1208–1216*

[3] Resultados de las estimaciones de ancestría con toda la muestra, su correlación entre las plataformas usadas y su asociación con neoplasias colorrectales, publicado en: *Sanabria-Salas MC, et al. IL1B-CGTC haplotype is associated with colorectal cancer in admixed individuals with increased African ancestry. Scientific Reports. 2017;7:41920, DOI: 10.1038/srep41920.*

Por su parte, la incidencia estimada anual de CCR en mujeres durante el periodo 2007 a 2011, fue mayor en Quindío, Risaralda, Caldas, Bogotá DC, Valle del Cauca, Meta, San Andrés y Providencia, Antioquia, Tolima y Santander con valores de TAE de 18.5 a 12.5; mientras que los departamentos con menor incidencia fueron Huila, Sucre, Caquetá, Cesar, Magdalena, Casanare, Arauca, Bolívar, Nariño, Cauca, Putumayo, Córdoba, Chocó, la Guajira y Grupo Amazonas, con TAE de 8.8 a 4.0 (ver **Tabla 5**) (111).

Tabla 4 Incidencia y mortalidad por CCR en hombres colombianos, según departamento

DEPARTAMENTOS	HOMBRES COLOMBIANOS					
	INCIDENCIA ESTIMADA ANUAL		MORTALIDAD OBSERVADA ANUAL			REM (2000-2006)
	TAE (2002-2006)	TAE (2007-2011)	TAE (2002-2006)	TAE (2007-2011)	CAMBIO RELATIVO	
Antioquia	14.1	13.5	6.3	6.7	6%	**117**
Arauca	5.0	10.4	2.1	4.7	124%	56
Atlántico*	10.7	9.1	4.6	4.5	-2%	88
Bogotá*	19.0	18.0	8.7	9.0	3%	**155**
Bolívar*	8.2	7.8	3.4	3.8	12%	70
Boyacá	10.4	10.8	4.6	5.6	22%	87
Caldas	15.7	15.9	6.8	7.6	12%	**134**
Caquetá	5.9	4.5	2.6	1.6	-38%	60
Casanare	9.8	8.0	3.9	3.6	-8%	71
Cauca	6.3	5.8	2.6	2.8	8%	51
Cesar	7.1	8.2	3.1	3.6	16%	58
Chocó	5.8	3.0	2.3	1.9	-17%	37
Córdoba	5.2	7.0	2.3	3.4	48%	41
Cundinamarca	12.1	13.5	5.4	6.6	22%	98
Huila	9.0	8.4	4.0	3.9	-3%	72
La Guajira	3.9	3.0	1.8	1.6	-11%	33
Magdalena*	7.9	7.3	3.5	3.6	3%	69
Meta	14.8	14.0	6.6	7.1	8%	**117**
Nariño	5.5	6.0	2.3	2.9	26%	43
Norte de Santander	10.7	9.5	4.7	4.8	2%	90
Putumayo	4.9	7.0	2.1	3.5	67%	48
Quindío	12.1	18.3	5.4	8.4	56%	**101**
Risaralda	15.7	16.0	7.1	7.7	8%	**134**
San Andrés y Providencia	12.5	7.6	4.1	4.3	5%	86
Santander*	13.0	13.8	5.8	6.8	17%	**111**
Sucre	6.1	6.4	2.7	2.8	4%	56
Tolima	8.7	11.9	3.9	5.8	49%	78
Valle del Cauca*	14.0	14.9	6.2	7.3	18%	**123**
Grupo Amazonas**	5.5	7.0	2.0	2.9	45%	52
Total Colombia	11.9	12.2	5.3	6.0	13%	100

* Departamentos de Colombia en cuyas capitales se realizaron las actividades de captura de casos y controles de este trabajo de tesis.

** Incluye Amazonas, Guainía, Guaviare, Vichada y Vaupés.

TAE, tasas ajustadas por edad por cada 100 mil habitantes. Datos tomados de *Pardo C et al., 2010; Pardo C et al., 2015* (111, 112).
CAMBIO RELATIVO, son las diferencias en el número de muertes observadas en los dos periodos mencionados en la tabla. Se calculó así: tasas de mortalidad del periodo 2007-2011 menos las tasas de mortalidad del periodo 2002-2006, dividido en las tasas de mortalidad del periodo 2002-2006 (expresado en porcentaje).
REM, son las razones de mortalidad estandarizadas. Los valores REM > 100 se resaltan en negrilla. Datos tomados *de Piñeros M et al* (2).

Tabla 5 Incidencia y mortalidad por CCR en mujeres colombianas, según departamento

DEPARTAMENTOS	MUJERES COLOMBIANAS					
	INCIDENCIA ESTIMADA ANUAL		MORTALIDAD OBSERVADA ANUAL			REM (2000-2006)
	TAE (2002-2006)	TAE (2007-2011)	TAE (2002-2006)	TAE (2007-2011)	CAMBIO RELATIVO	
Antioquia	14.9	13.5	6.9	6.7	-3%	**123**
Arauca	6.7	8.4	2.9	3.8	31%	52
Atlántico*	12.2	11.9	5.8	6.0	3%	98
Bogotá*	15.2	15.9	7.1	7.8	10%	**127**
Bolívar*	9.8	8.1	4.5	4.0	-11%	79
Boyacá	10.2	10.4	4.7	5.1	9%	74
Caldas	17.2	16.2	8.0	8.0	0%	**137**
Caquetá	8.0	8.7	3.4	3.7	9%	58
Casanare	5.6	8.5	2.8	3.6	29%	45
Cauca	6.9	6.8	3.3	3.3	0%	63
Cesar	7.4	8.7	3.2	4.0	25%	71
Chocó	5.2	5.5	2.1	1.8	-14%	30
Córdoba	6.8	5.6	3.2	2.7	-16%	52
Cundinamarca	10.4	9.9	4.7	4.9	4%	82
Huila	7.8	8.8	3.7	4.5	22%	67
La Guajira	5.3	4.5	2.3	1.8	-22%	41
Magdalena*	8.5	8.7	4.0	4.5	13%	72
Meta	10.0	15.0	4.6	7.6	65%	90
Nariño	6.9	7.2	3.1	3.5	13%	53
Norte de Santander	13.6	12.0	6.4	6.0	-6%	**110**
Putumayo	3.6	6.5	1.6	2.9	81%	29
Quindío	14.0	18.5	6.5	8.5	31%	**116**
Risaralda	17.9	16.6	8.2	8.1	-1%	**143**
San Andrés y Providencia	4.2	13.7	3.4	7.1	109%	40
Santander*	13.1	12.5	5.9	6.1	3%	**102**
Sucre	7.4	8.8	3.6	4.2	17%	64
Tolima	11.1	13.3	5.2	6.7	29%	91
Valle del Cauca*	15.0	15.3	7.0	7.5	7%	**122**
Grupo Amazonas**	3.4	4.0	1.0	1.4	40%	30
Total Colombia	12.3	12.3	5.7	6.1	7%	100

* Departamentos de Colombia en cuyas capitales se realizaron las actividades de captura de casos y controles de este trabajo de tesis.
** Incluye Amazonas, Guainía, Guaviare, Vichada y Vaupés.
TAE, tasas ajustadas por edad por cada 100 mil habitantes. Datos tomados de *Pardo C et al., 2010; Pardo C et al., 2015* (111, 112).

CAMBIO RELATIVO, son las diferencias en el número de muertes observadas en los dos periodos mencionados en la tabla. Se calculó así: tasas de mortalidad del periodo 2007-2011 menos las tasas de mortalidad del periodo 2002-2006, dividido en las tasas de mortalidad del periodo 2002-2006 (expresado en porcentaje).
REM, son las razones de mortalidad estandarizadas. Los valores REM > 100 se resaltan en negrilla. Datos tomados *de Piñeros M et al* (2).

Por otro lado, los departamentos con mayor mortalidad observada anual en hombres, durante el periodo 2007 a 2011, fueron Bogotá DC, Quindío, Risaralda, Caldas, Valle del Cauca, Meta, Santander, Antioquia, Cundinamarca y Tolima, con valores de TAE de 9.0 a 5.8; mientras que los departamentos con menores tasas de mortalidad fueron Magdalena, Putumayo, Córdoba, Nariño, Grupo Amazonas, Cauca, Sucre, Chocó, Caquetá y la Guajira, con TAE de 3.6 a 1.6 (ver **Tabla 4**) (111).

Con respecto a la mortalidad observada anual de CCR en mujeres, fue mayor en el Quindío, Risaralda, Caldas, Bogotá DC, Meta, Valle del Cauca, San Andrés y Providencia, Antioquia, Tolima y Santander, con valores TAE de 8.5 a 6.1; mientras que los departamentos con menores tasas de mortalidad fueron Arauca, Caquetá, Casanare, Nariño, Cauca, Putumayo, Córdoba, Chocó, la Guajira y Grupo Amazonas, con TAE de 3.8 a 1.4 (ver **Tabla 5**) (111).

Teniendo en cuenta que las tasas de mortalidad por CCR fueron calculadas a partir de los datos observados para los dos periodos, 2002-2006 y 2007-2011, la metodología no cambio y es posible compararlos; para esto, se calculó el cambio relativo de las tasas de mortalidad y se describen en porcentajes (ver **Tabla 4** y **Tabla 5**). Los departamentos en los cuales hubo un mayor incremento relativo en las muertes en hombres fueron en Arauca, Putumayo, Quindío, Tolima, Córdoba y Grupo Amazonas, con valores que van desde un 124% a 45% más de casos en 2007-2011 con respecto al periodo 2002-2006 (ver **Tabla 4**). En los departamentos del Nariño, Cundinamarca, Boyacá, Valle del Cauca, Santander, Cesar, Caldas y Bolívar, este incremento estuvo entre un 26% a 12% y en Risaralda, Cauca, Meta, Antioquia, San Andrés y Providencia, Sucre, Bogotá DC, Magdalena y Norte de Santander, el cambio fue de un 8% al 2% en el número de muertes observadas entre los dos periodos mencionados (ver **Tabla 4**). De otro lado, en los departamentos del Atlántico, Huila, Casanare, la Guajira, Chocó y Caquetá se observó una disminución de las muertes en el 2007-2011 con respecto al periodo 2002-2006, desde un -2% hasta un -38% (ver **Tabla 4**).

En mujeres igualmente se observaron diferencias en el número de muertes por CCR entre los dos periodos. En los departamentos de San Andrés y Providencia, Putumayo, Meta, Grupo Amazonas, Arauca, Quindío, Tolima, Casanare y Cesar, hubo un aumento relativo del 109% al 25% (ver **Tabla 5**). En Huila, Sucre, Nariño, Magdalena y Bogotá DC, el cambio fue de un 22% a 10% y en Caquetá, Boyacá, Valle del Cauca, Cundinamarca, Atlántico y Santander, este incremento estuvo entre el 9% al 3% (ver **Tabla 5**). Por otro lado, se observó una disminución de las muertes en el periodo 2007-2011, con respecto al 2002-2006, en Risaralda, Antioquia, Norte de Santander, Bolívar, Chocó, Córdoba y la Guajira con valores de -1% al -22%; mientras que en Caldas y Cauca no se observó ningún cambio (ver **Tabla 5**).

De acuerdo a las razones de mortalidad estandarizadas (REM) (reportadas para el periodo entre el 2000 al 2006), que habla del riesgo de muerte por una enfermedad en una población determinada con respecto a la población general, se observó una concentración del riesgo de muerte mayor que el promedio nacional en ciudades del centro del país a lo largo de la cordillera central, incluyendo desde Cali, en el Valle del Cauca, hasta Medellín en Antioquía (2). Otros focos de mayor riesgo se ubicaron en la cordillera occidental, en Bogotá DC y los Santanderes; mientras que las zonas de bajo riesgo correspondieron a los departamentos de Nariño, Cauca, la Guajira y Magdalena Medio al norte (2). En los departamentos de Bogotá DC, Valle del Cauca, Caldas y Risaralda, se evidenció un aumento en el riesgo de muerte con respecto a la población general por encima del 22% en ambos sexos (REM ≥ 122) (ver **Tabla 4** y **Tabla 5**) (2).

Según lo reportado por *Piñeros M, et al.*, en un análisis de tendencia en el periodo 1985-2006, la mortalidad por CCR en Colombia mostró un incremento anual promedio del 2.2% y 1.9%, entre hombres y mujeres, respectivamente ($P < 0.001$) (2). Particularmente en hombres, este incremento anual promedio fue significativo en Bogotá DC, Antioquia, Valle del Cauca y Santander, siendo mayor en estos dos últimos con un 2.2% y 2.3%, respectivamente ($P < 0.01$) (2). En mujeres, el incremento anual promedio en la mortalidad por CCR fue significativo en Valle del Cauca, Santander, Atlántico y Boyacá, con valores desde 2.2% hasta 2.5%, siendo mayor en Atlántico y Boyacá ($P < 0.01$) (2). Llama la atención que la tendencia al aumento en la mortalidad por CCR en Santander se

observó en ambos sexos con valores de 2.3% y 2.4% en hombres y mujeres, respectivamente, y en Valle del Cauca con 2.2% en ambos sexos (2).

1.1 Estudio de la variabilidad en la estructura genética y de sesgos de género en el proceso de mestizaje en poblaciones colombianas con diferente riesgo de CCR, mediante el análisis de las proporciones de ancestrías europea, amerindia y africana:

Teniendo en cuenta la gran diversidad en los componentes ancestrales de la población colombiana, dadas por las características geográficas, migratorias y de mestizaje a lo largo del territorio, seleccionamos para este estudio seis (6) capitales departamentales de las regiones Andina (Bogotá DC y Bucaramanga) y Costera (Cali en el Pacífico y, Barranquilla, Cartagena y Santa Marta en el Caribe) del país. Igualmente, para esta selección se tuvieron en cuenta las diferencias en el riesgo de CCR descritas previamente en este capítulo, además de otros aspectos relacionados con la logística del estudio.

Como se observa en la **Figura 3**, y de acuerdo a lo descrito anteriormente, en las poblaciones de Bogotá DC, Valle del Cauca (capital Cali) y Santander (capital Bucaramanga) se han evidenciado altas tasas de incidencia y mortalidad por CCR, a su vez que ha habido un incremento significativo en el promedio anual en las tasas de mortalidad, especialmente en Valle del Cauca y Santander; más aún, según el *Atlas de mortalidad por cáncer en Colombia 2000-2006*, estas tres poblaciones mostraron un mayor riesgo de mortalidad por CCR que la población general colombiana (2, 111).

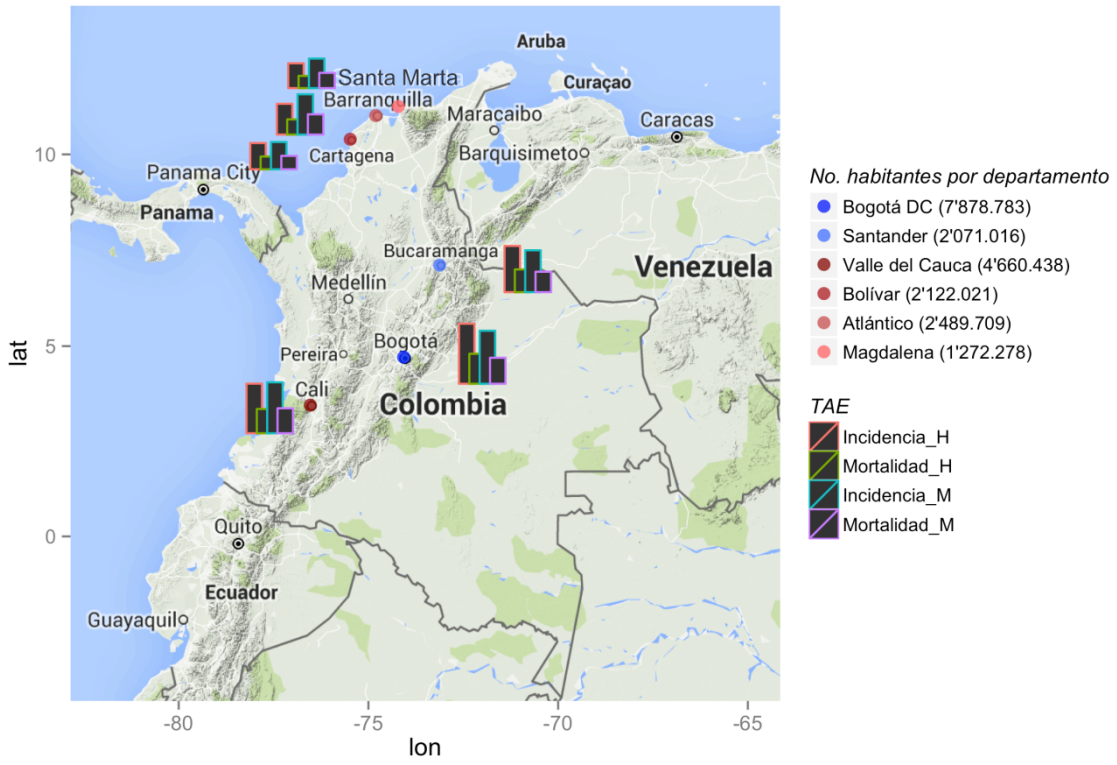


Figura 3 Diferencias en las tasas de incidencia y mortalidad ajustadas por edad (TAE) en CCR, discriminadas por sexo, entre departamentos de las regiones andina y costera de Colombia. Solo se muestra la información de los seis (6) departamentos cuyas capitales fueron sitios de captura de casos y controles incluidos en este trabajo de tesis doctoral. Los datos de incidencia y mortalidad corresponden al grupo conformado por cáncer de colon + cáncer de recto + cáncer de ano, tal y como se describe en *Pardo C et al* (111), los cuales no fue posible reportar por separado. Los puntos de color azul corresponden a ciudades de los Andes y los puntos de color rojo a ciudades de la Costa Caribe y Pacífica. Figura original generada en el programa *R statistics* (76).

TAE, tasas ajustadas por edad por cada 100 mil habitantes; Incidencia_H, incidencia en hombres; Mortalidad_H, mortalidad en hombres; Incidencia_M, incidencia en mujeres; Mortalidad_M, mortalidad en mujeres; lon, longitud; lat, latitud; Bogotá DC, capital de Colombia; Santander, capital Bucaramanga; Valle del Cauca, capital Cali; Bolívar, capital Cartagena; Atlántico, capital Barranquilla; Magdalena, capital Santa Marta.

El propósito de esta primera parte del estudio, es medir objetivamente las diferencias en las proporciones de ancestría en las poblaciones incluidas y evaluar posibles sesgos de género que hicieron parte del proceso de mestizaje.

1.1.1 Objetivos específicos

1.1.1.1 Estimar las proporciones de ancestría globales a nivel de genoma completo, autosomas y cromosoma X, en colombianos incluidos en el estudio.

1.1.1.2 Evaluar diferencias en las estimaciones en cromosoma X versus autosomas para la población colombiana, y discriminado por regiones y por ciudades

1.1.1.3 Calcular la proporción de mujeres de ancestría europea, amerindia y africana, que contribuyeron al proceso de mestizaje de las poblaciones incluidas en el estudio

1.1.2 Métodos

Para alcanzar los tres objetivos planteados en esta primera parte del estudio, se incluyeron las muestras genotipadas con la plataforma o microarreglo de genoma completo (GWAS), que pasaron el control de calidad por individuos, n = 415 (ver **Tabla - Anexos A**). Solo se usó ésta base GWAS porque contiene el suficiente número de SNPs en el cromosoma X necesarios para obtener estimaciones precisas y compararlas con las obtenidas en autosomas. *Ver el Anexo 1 - Materiales y métodos en análisis genéticos, para obtener información más detallada de los métodos.*

Las estimaciones de ancestría se calcularon usando como referencia las poblaciones europeas [n = 107; IBS = población de España] y africanas [n = 108; YRI = africanos de Ibadan, Nigeria] de la base pública de *1000 Genomes* (113) y las poblaciones amerindias [n = 108; poblaciones de nativos americanos, AME= Pima, Maya, Karitiana, Surui y de Colombia] de la base pública del *HGDP* (114). La base de datos consolidada (415 muestras colombianas + 323 de referencia = 738), incluyó los **genotipos** de 9868 SNPs comunes entre todas las bases de datos (9663 en autosomas + 205 en cromosoma X),

despues de los pasos de control de calidad por SNPs y de reducción de la redundancia de SNPs (ver **Figura - Anexos F**).

Para estimar las proporciones de ancestría global europea, amerindia y africana en las muestras colombianas, se usó el algoritmo de ADMIXTURE v1.3.0 (115), fijando el número de componentes ancestrales a k = 3. Se calcularon las estimaciones en tres escenarios por separado: a nivel de genoma-completo (autosomas + cromosoma X), en autosomas y en cromosoma X. Teniendo en cuenta que los hombres son **hemicigotos** para el cromosoma X, siempre que se incluyeron marcadores en éste cromosoma, se añadió el comando --haploid="male:23" (116). Una vez obtenidas todas las estimaciones, se seleccionaron solo las muestras *"controles"* (n = 131; mujeres 73 y hombres 58), para hacer las respectivas comparaciones.

Para evaluar diferencias en las ancestrías calculadas en autosomas versus cromosoma X, se usó el test no paramétrico Wilcoxon-Mann-Whitney y un test basado en 100 mil permutaciones (116). También se calculó la razón normalizada de las diferencias en las ancestrías calculadas en el cromosoma X versus las calculadas en autosomas, como se describe en la **Tabla 7**, **Tabla 8** y **Tabla 9**, y de acuerdo a *Rishishwar L, et al* (9). Por otro lado, se calculó la proporción de hombres y mujeres de cada ancestría (europea, amerindia y africana), que participaron en el proceso de mestizaje de las ciudades y regiones de Colombia incluidas; para lo anterior, se reemplazaron las medias de las proporciones de ancestría observadas en *"controles"* a nivel de autosomas ($H^{A}_{1,g,\delta}$) y de cromosoma X ($H^{X}_{1,g,\delta}$) en las ecuaciones (a) y (b), que además están descritas en la publicación de *Goldberg A, et al* (87). Para estos últimos cálculos, se asumió una población con igual número de hombres que de mujeres, que resultó de un (1) evento de mestizaje en un momento en el tiempo.

1.1.3 Resultados

Estimación de las proporciones de ancestría globales a nivel de genoma completo, autosomas y cromosoma X en "controles" colombianos incluidos en el estudio:

En el total de la muestra *"control"* analizada, se obtuvieron medias de ancestrías europea, amerindia y africana de 54%, 34% y 11%, respectivamente; mientras que las medias en cromosoma X para estos tres componentes fueron 31%, 54% y 14% (ver **Tabla 6** y **Figura 4**). Al comparar estos estimados, se encontraron diferencias significativas tanto en las ancestrías europea y amerindia ($P < 0.001$; $P_{100\ mil\ permutaciones} < 0.001$), como en la africana ($P < 0.08$; $P_{100\ mil\ permutaciones} = 0.05$). En general, se observa un exceso en la ancestría europea en autosomas en comparación con el cromosoma X; mientras que en el cromosoma X la proporción de ancestría amerindia fue mayor respecto a la media en autosomas (ver **Tabla 6** y **Figura 4**).

Tabla 6 Distribución de las proporciones de ancestrías estimadas en *"controles"* colombianos, usando marcadores biparentales

DISTRIBUCIONES	PROPORCIONES DE ANCESTRÍAS EN CONTROLES COLOMBIANOS					
	AUTOSOMAS			CROMOSOMA X		
	Europea	Amerindia	Africana	Europea	Amerindia	Africana
Mínimo	0.05	0.08	0.00	0.00	0.00	0.00
1er cuartil	0.48	0.27	0.01	0.03	0.35	0.00
Mediana	0.57	0.35	0.04	0.27	0.50	0.03
Media [95% CI]	0.54 [0.52-0.57]	0.34 [0.33-0.36]	0.11 [0.09-0.14]	0.31 [0.26-0.36]	0.54 [0.49-0.60]	0.14 [0.11-0.18]
3er cuartil	0.63	0.40	0.18	0.52	0.84	0.22
Máximo	0.84	0.72	0.87	0.92	1.00	1.00

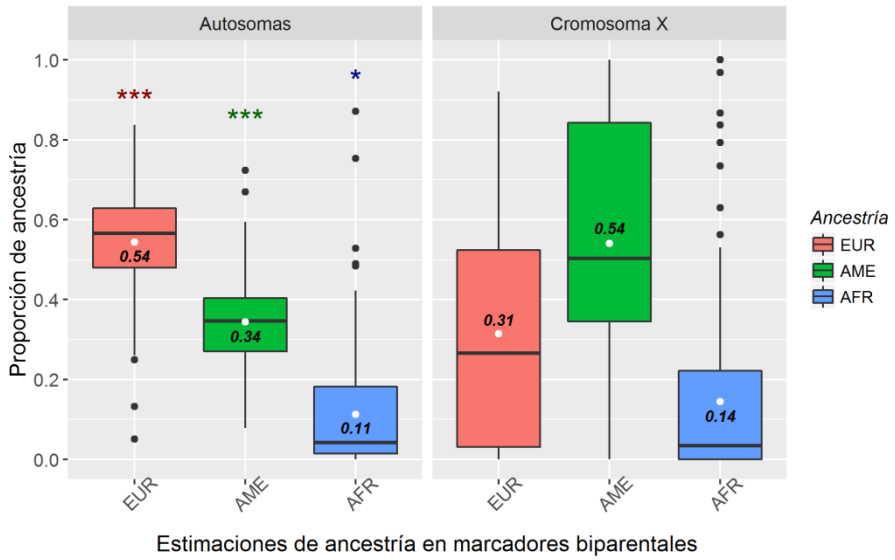


Figura 4 Comparación de la distribución de las proporciones de ancestrías estimadas en *"controles"* colombianos, usando marcadores biparentales. Gráfico tipo cajas y bigotes que permite observar la distribución de las proporciones de cada ancestría [EUR = europeo; AME = amerindio; AFR = africano] en autosomas y en cromosoma X (ver **Tabla 6**). Los puntos blancos indican la *media* de la distribución (con sus respectivos valores anotados). Figura original generada en el programa *R statistics* (76).
*** Valores de *P* significativos ($P_{\text{100 mil permutaciones}} < 0.001$), calculados con un test de permutaciones, al comparar las proporciones de ancestrías europea y amerindia obtenidas en autosomas versus cromosoma X.
* Valores de *P* significativos ($P_{\text{100 mil permutaciones}} = 0.05$), calculados con un test de permutaciones, al comparar las proporciones de ancestría africana, obtenidas en autosomas versus cromosoma X.
EUR, europeo (rojo); AME, amerindio (verde); AFR, africano (azul)

Comparación de las estimaciones de ancestrías en cromosoma X versus autosomas para la población colombiana, discriminado por regiones y por ciudades:

En las comparaciones discriminadas por regiones y por ciudades, se observan igualmente diferencias importantes en las proporciones de ancestría en autosomas versus cromosoma X (ver **Figura 5** y **Figura 6**, respectivamente); mientras que las estimaciones obtenidas a nivel de genoma completo y autosomas fueron casi iguales (ver **Tabla 7**, **Tabla 8** y **Tabla 9**).

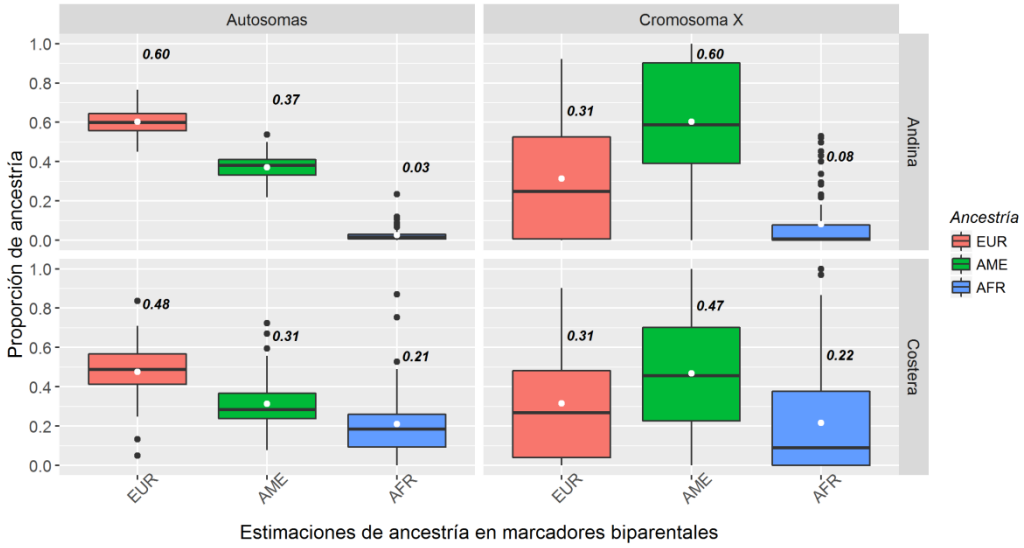


Figura 5 Distribución de las proporciones de ancestrías estimadas en *"controles"* colombianos, usando marcadores biparentales, discriminado por regiones, andina y costera. Gráfico tipo cajas y bigotes que permite observar la distribución de las proporciones de cada ancestría en autosomas y en cromosoma X (ver **Tabla 7**). Los puntos blancos indican la *media* de la distribución (con sus respectivos valores anotados). Figura original generada en el programa *R statistics* (76).
EUR, europeo (rojo); AME, amerindio (verde); AFR, africano (azul)

Tabla 7 Sesgos de género en el proceso de mestizaje en la región andina y costera de Colombia, usando marcadores biparentales

	ANCESTRÍAS		
ESTIMACIONES §	**Europea**	**Amerindia**	**Africana**
Región Andina (n total = 70)			
Genoma completo	60%	38%	3%
Autosomas	60%	37%	3%
Cromosoma X	31%	60%	8%
*Valor de P (P 100 mil perm)**	*1.23×10^{-9} (2.00×10^{-5})*	*2.24×10^{-7} (2.00×10^{-5})*	*0.71 (6.40×10^{-4})*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-19%	9%	1%
% de mujeres***	22%	77%	58%
Región Costera (n total = 61)			
Genoma completo	47%	32%	21%
Autosomas	48%	31%	21%
Cromosoma X	31%	47%	22%
*Valor de P (P 100 mil perm)**	*1.73×10^{-4} (2.00×10^{-5})*	*1.45×10^{-3} (6.00×10^{-5})*	*0.06 (0.84)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-10%	6%	0%
% de mujeres***	1%	84%	54%
Diferencias entre regiones****			
Autosomas	*1.04×10^{-8} (2.00×10^{-5})*	*6.26×10^{-5} (1.28×10^{-3})*	*1.08×10^{-15} (2.00×10^{-5})*
Cromosoma X	*0.79 (0.99)*	*0.02 (0.01)*	*3.45×10^{-3} (3.80×10^{-4})*

§ Estimaciones observadas de ancestrías europea, amerindia y africana en genoma completo (autosomas + cromosoma X), en autosomas (cromosoma 1 a 22) y cromosoma X (cromosoma 23). Los análisis se corrieron en ADMIXTURE y cuando se analizó el cromosoma X se especificó --haploid="male:23", debido a que los hombres son hemicigotos para este cromosoma.

* Valor de *P* obtenido con el test de Wilcoxon-Mann-Whitney (valor de *P* obtenido con un test basado en 100 mil permutaciones), para evaluar diferencias significativas entre cada componente de ancestría estimado en autosomas versus cromosoma X.

** La razón normalizada de las diferencias en las ancestrías estimadas en cromosoma X - autosomas se calculó así: $(\overline{\Delta Admix})$ = $GC_{anc} * (CrX_{anc} - Auto_{anc})/(CrX_{anc} + Auto_{anc})$; donde GC_{anc}, CrX_{anc} y $Auto_{anc}$ son las proporciones de cada ancestría a nivel de genoma completo, en el cromosoma X y en autosomas, respectivamente, según *Rishishwar L et al* (9).

*** El porcentaje (%) de mujeres, se calculó así: $\%_{muj} = \%_{relativo,muj} / (\%_{relativo,muj} + \%_{relativo,hom})$; donde el $\%_{relativo,muj} = 3CrX_{anc} - 2Auto_{anc}$ y el $\%_{relativo,hom} = 2Auto_{anc} - \%_{relativo,muj}$. Lo anterior, de acuerdo a la Ecuación (a) y (b), descrita en la publicación de *Goldberg A et al* (87) y en este capítulo.

En el análisis discriminado por regiones, en ambas se observó la misma tendencia que en el análisis de toda la muestra; es decir, una predominancia de las ancestrías europea en autosomas y amerindia en cromosoma X ($P_{\text{100 mil permutaciones}} < 0.001$); mientras que solo en la region andina se encontró un incremento en el componente africano a nivel del

cromosoma X, comparado con autosomas, el cual fue estidisticamente significativo ($P_{100\ mil\ permutaciones} < 0.001$) (ver **Tabla 7** y **Figura 5**). La ancestría europea en el cromosoma X fue 19% y 10% menos que en los autosomas en las regiones andina y costera, respectivamente; mientras que la ancestría amerindia fue 9% y 6% mayor en el cromosoma X que en autosomas, en estas mismas regiones (ver **Tabla 7**). Al comparar las proporciones de cada ancestría en autosomas entre las regiones, andina versus costera, se encontraron diferencias significativas en todos los componentes ($P_{100\ mil\ permutaciones} < 0.001$), principalmente en el europeo (60% versus 48%, respectivamente) y en el africano (3% versus 21%, respectivamente); por otro lado, al hacer esta misma comparación con las proporciones de ancestría en cromosoma X, se encontraron diferencias principalmente para la ancestría africana (8% versus 22%; $P_{100\ mil\ permutaciones} \leq 0.001$) y en menor medida para la ancestría amerindia (60% versus 47%; $P_{100\ mil\ permutaciones} = 0.01$), pero no para la ancestría europea (ver **Tabla 7**).

En el análisis discriminado por ciudades de la región andina, Bogotá y Bucaramanga, se observó un predominio del 19% en la ancestría europea en autosomas y del 9% en la ancestría amerindia en cromosoma X en ambos casos ($P_{100\ mil\ permutaciones} < 0.001$); igualmente, se evidencia un aumento de solo el 1% en el componente africano a nivel del cromosoma X versus autosomas ($P_{100\ mil\ permutaciones} < 0.05$) (ver **Tabla 8** y **Figura 6**). Con respecto a las ciudades de la región costera (Cali en el Pacífico y Barranquilla, Cartagena y Santa Marta en el Caribe), el predominio de la ancestría europea en autosomas fue más marcado en las tres (3) ciudades de la Costa Caribe (~11%) que en la Costa Pacífica (7%), aunque solo fue significativo en Barranquilla, Cartagena y Cali ($P_{100\ mil\ permutaciones} < 0.05$); igualmente, el predominio de la ancestría amerindia en cromosoma X fue mayor en las ciudades del Caribe (~7%) que en la Costa Pacífica (~3%), siendo significativo solo en Barranquilla y Cartagena ($P_{100\ mil\ permutaciones} < 0.05$) (ver **Tabla 9** y **Figura 6**). Con respecto a las estimaciones del componente africano, no se observaron diferencias significativas en ninguna ciudad costera (ver **Tabla 9**).

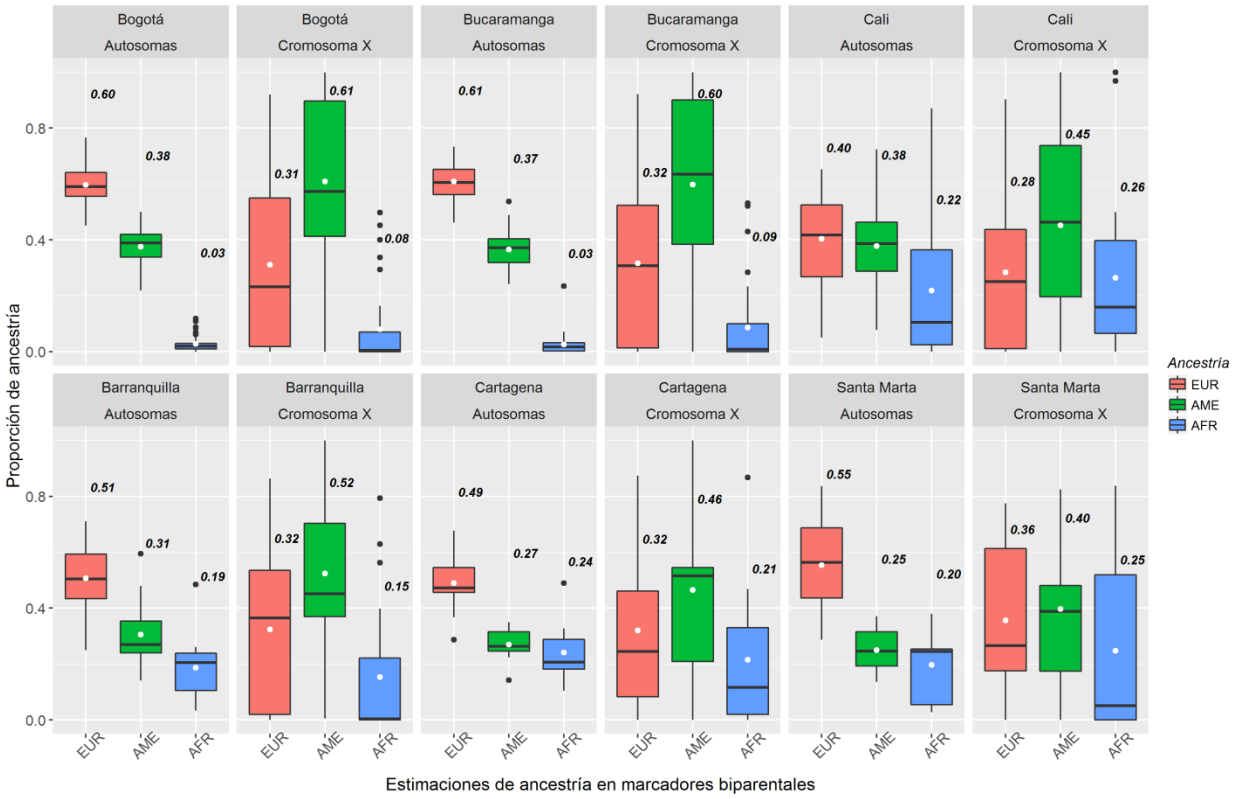


Figura 6 Distribución de las proporciones de ancestrías estimadas en *"controles"* colombianos, usando marcadores biparentales, discriminado por ciudades de las regiones andina (Bogotá y Bucaramanga) y costera (Cali en el Pacífico y Barranquilla, Cartagena y Santa Marta en el Caribe). Gráfico tipo cajas y bigotes que permite observar la distribución de las proporciones de cada ancestría en autosomas y en cromosoma X (ver **Tabla 8** y **Tabla 9**). Los puntos blancos indican la *media* de la distribución (con sus respectivos valores anotados). Figura original generada en el programa *R statistics* (76).
EUR, europeo (rojo); AME, amerindio (verde); AFR, africano (azul)

Tabla 8 Sesgos de género en el proceso de mestizaje en las ciudades de la región andina de Colombia, usando marcadores biparentales

	ANCESTRÍAS		
ESTIMACIONES§	**Europea**	**Amerindia**	**Africana**
Región Andina (n total = 70)			
Bogotá (n = 35)			
Genoma completo	59%	38%	3%
Autosomas	60%	38%	3%
Cromosoma X	31%	61%	8%
*Valor de P (P 100 mil perm)**	*1.29×10^{-5} (2.00×10^{-5})*	*1.28×10^{-4} (1.00×10^{-4})*	*0.54 (0.04)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-19%	9%	1%
% de mujeres***	22%	77%	59%
Bucaramanga (n = 35)			
Genoma completo	60%	37%	3%
Autosomas	61%	37%	3%
Cromosoma X	32%	60%	9%
*Valor de P (P 100 mil perm)**	*2.19×10^{-5} (2.00×10^{-5})*	*4.44×10^{-4} (1.80×10^{-4})*	*1.00 (0.01)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-19%	9%	1%
% de mujeres***	22%	76%	57%

§ Estimaciones observadas de ancestrías europea, amerindia y africana en genoma completo (autosomas + cromosoma X), en autosomas (cromosoma 1 a 22) y cromosoma X (cromosoma 23). Los análisis se corrieron en ADMIXTURE y cuando se analizó el cromosoma X se especificó --haploid="male:23", debido a que los hombres son hemicigotos para este cromosoma.

* Valor de *P* obtenido con el test de Wilcoxon-Mann-Whitney (valor de *P* obtenido con un test basado en 100 mil permutaciones), para evaluar diferencias significativas entre cada componente de ancestría estimado en autosomas versus cromosoma X.

** La razón normalizada de las diferencias en las ancestrías estimadas en cromosoma X - autosomas se calculó así: $(\overline{\Delta Admix})$ = $GC_{anc} * (CrX_{anc} - Auto_{anc})/(CrX_{anc} + Auto_{anc})$; donde GC_{anc}, CrX_{anc} y $Auto_{anc}$ son las proporciones de cada ancestría a nivel de genoma completo, en el cromosoma X y en autosomas, respectivamente, según *Rishishwar L et al* (9).

*** El porcentaje (%) de mujeres, se calculó así: $\%_{muj} = \%_{relativo,muj} / (\%_{relativo,muj} + \%_{relativo,hom})$; donde el $\%_{relativo,muj} = 3CrX_{anc} - 2Auto_{anc}$ y el $\%_{relativo,hom} = 2Auto_{anc} - \%_{relativo,muj}$. Lo anterior, de acuerdo a la Ecuación (a) y (b), descrita en la publicación de *Goldberg A et al* (87) y en este capítulo.

Tabla 9 Sesgos de género en el proceso de mestizaje en las ciudades de la región costera de Colombia, usando marcadores biparentales

	ANCESTRÍAS		
ESTIMACIONES§	**Europea**	**Amerindia**	**Africana**
Región Costera (n total = 61)			
Costa Pacífica			
Cali (n = 20)			
Genoma completo	40%	38%	22%
Autosomas	40%	38%	22%
Cromosoma X	28%	45%	26%
*Valor de P (P 100 mil perm)**	*0.06 (0.04)*	*0.49 (0.17)*	*0.64 (0.15)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-7%	3%	2%
% de mujeres***	5%	80%	81%
Costa Caribe			
Barranquilla (n = 19)			
Genoma completo	50%	31%	19%
Autosomas	51%	31%	19%
Cromosoma X	32%	52%	15%
*Valor de P (P 100 mil perm)**	*0.04 (0.01)*	*0.01 (0.01)*	*0.02 (0.61)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-11%	8%	-2%
% de mujeres***	4%	73%	22%
Cartagena (n = 13)			
Genoma completo	49%	27%	24%
Autosomas	49%	27%	24%
Cromosoma X	32%	46%	21%
*Valor de P (P 100 mil perm)**	*0.03 (0.03)*	*0.09 (0.04)*	*0.30 (0.69)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-10%	7%	-1%
% de mujeres***	2%	73%	33%
Santa Marta (n = 9)			
Genoma completo	55%	25%	20%
Autosomas	55%	25%	20%
Cromosoma X	36%	40%	25%
*Valor de P (P 100 mil perm)**	*0.14 (0.10)*	*0.19 (0.15)*	*0.48 (0.60)*
Diferencias en ancestrías de cromosoma X vs autosomas (razón normalizada) **	-12%	6%	2%
% de mujeres***	4%	79%	89%

§ Estimaciones observadas de ancestrías europea, amerindia y africana en genoma completo (autosomas + cromosoma X), en autosomas (cromosoma 1 a 22) y cromosoma X (cromosoma 23). Los análisis se corrieron en ADMIXTURE y cuando se analizó el cromosoma X se especificó --haploid="male:23", debido a que los hombres son hemicigotos para este cromosoma.

* Valor de *P* obtenido con el test de Wilcoxon-Mann-Whitney (valor de *P* obtenido con un test basado en 100 mil permutaciones), para evaluar diferencias significativas entre cada componente de ancestría estimado en autosomas versus cromosoma X.
** La razón normalizada de las diferencias en las ancestrías estimadas en cromosoma X - autosomas se calculó así: $(\overline{\Delta Admix})$ = GC_{anc} * $(CrX_{anc} - Auto_{anc})/(CrX_{anc} + Auto_{anc})$; donde GC_{anc}, CrX_{anc} y $Auto_{anc}$ son las proporciones de cada ancestría a nivel de genoma completo, en el cromosoma X y en autosomas, respectivamente, según *Rishishwar L et al* (9).
*** El porcentaje (%) de mujeres, se calculó así: $\%_{muj} = \%_{relativo,muj} / (\%_{relativo,muj} + \%_{relativo,hom})$; donde el $\%_{relativo,muj} = 3CrX_{anc} - 2Auto_{anc}$ y el $\%_{relativo,hom} = 2Auto_{anc} - \%_{relativo,muj}$. Lo anterior, de acuerdo a la Ecuación (a) y (b), descrita en la publicación de *Goldberg A et al* (87) y en este capítulo.

Proporción de mujeres de ancestría europea, amerindia y africana, que contribuyeron al proceso de mestizaje en ciudades colombianas (sesgos de género):

El cálculo de la proporción de mujeres de cada población ancestral que contribuyeron al proceso de mestizaje, se realizó por regiones y por ciudades colombianas, y permitió evidenciar una contribución asimétrica de hombres y mujeres de origen europeo, amerindio y africano en este proceso. Los resultados fueron muy similares entre las ciudades de la región andina pero variables entre las ciudades de la región costera (ver **Tabla 7**, **Tabla 8** y **Tabla 9**). En general, tanto en Bogotá como en Bucaramanga se obtuvo que un ~77% de la **población efectiva** amerindia eran mujeres, mientras que un ~78% de la población efectiva europea eran hombres; igualmente, del 100% de individuos **fundadores** de África, hubo una ligera mayor contribución de mujeres (~58%), en relación a hombres en estas dos ciudades (ver **Tabla 7** y **Tabla 8**). Con respecto a las ciudades costeras, en Cali se obtuvo que solo el 5% de la población efectiva europea eran mujeres y, por consiguiente, el 95% eran hombres, similar a lo obtenido para Barranquilla y Santa Marta; mientras que en Cartagena la contribución de hombres europeos al proceso de mestizaje fue ligeramente mayor (98%) que en las demás ciudades de la costa (ver **Tabla 9**). En Cali se obtuvo que el 80% de la población efectiva amerindia eran mujeres, similar a Santa Marta; mientras que en Barranquilla y Cartagena este porcentaje fue un poco menor (73%) (ver **Tabla 9**). La mayor asimetría de género en la población efectiva africana se observó en Santa Marta y Cali, con un 89% y 81% de mujeres, respectivamente; por otro lado, esta asimetría fue un poco menor en Barranquilla y Cartagena, donde además se observó una mayor contribución de hombres africanos en el proceso de mestizaje, con un 78% y 67%, respectivamente (ver **Tabla 9**).

1.2 Estudio del efecto de la estructura genética de diferentes poblaciones colombianas sobre la variabilidad observada en el riesgo de tumores colorrectales en el país

En la primera parte de este capítulo, se expuso la evidencia científica encontrada sobre el efecto que tuvo el patrón asimétrico de mestizaje en la heterogeneidad observada de las proporciones de ancestría europea, amerindia y africana de las poblaciones colombianas estudiadas en este trabajo de tesis. Igualmente, se analizó la epidemiología del CCR en nuestro país, evidenciando la variabilidad en el riesgo de la enfermedad en Colombia.

Teniendo en cuenta la evidencia previamente expuesta sobre el rol de la ancestría genética en el riesgo de varios tipos de cáncer (88-92, 96, 99-101), se planteó como ***propósito de esta segunda parte del estudio, evaluar la influencia de la composición genética ancestral heterogénea de los colombianos en la modificación del riesgo de desarrollar PA y CCR en el país.***

1.2.1 Objetivos específicos

1.2.1.1 Describir las características de la totalidad de la muestra de casos y controles colombianos incluidos en el estudio

1.2.1.2 Evaluar diferencias en las proporciones de ancestría globales a nivel de autosomas entre casos y controles colombianos

1.2.1.3 Evaluar el efecto de factores no genéticos en el riesgo de PA y CCR en colombianos con diferentes proporciones de ancestría

1.2.2 Métodos

En este trabajo de tesis se incluyeron 506 controles, 200 casos de PA y 313 casos de CCR, para un total de 1019 colombianos originarios de las seis (6) ciudades seleccionadas. Se realizó un análisis descriptivo de las variables edad, sexo, nivel educativo, historia familiar de CCR, consumo de AINES y región de origen, para evaluar diferencias entre los grupos, usando el estadístico X^2 de Pearson. *Ver el Anexo 1 - Materiales y métodos en análisis genéticos, para obtener información más detallada de los métodos.*

Un aspecto muy importante en los estudios genéticos de casos y controles es evaluar si éstos son comparables y si provienen de la misma población expuesta a factores de riesgo similares, es decir, evaluar si existe o no estratificación poblacional o sesgos en la recolección de las muestras; esto se realizó usando el método de **escalamiento multidimensional (MDS)**. Una vez verificado el paso anterior, se realizaron las estimaciones de las proporciones de ancestría global europea, amerindia y africana en las muestras colombianas. Para estos análisis se usaron los genotipos en autosomas de 813 individuos únicos de Colombia, obtenidos mediante las siguientes metodologías:

- **Genotipificación mediante un microarreglo de genes candidatos (CG) conocido como *"Cancer SNP Panel".*** Esta base consiste en datos de 1237 SNPs en 483 muestras colombianas (ver **Figura - Anexos B** y **Tabla - Anexos A**); lo anterior después de aplicar los controles de calidad según el protocolo de *Anderson C, et al* (117). La base de datos consolidada y preparada para evaluar la estratificación poblacional, y estimar la ancestría global, incluyó los genotipos de 473 SNPs comunes entre las muestras colombianas y las poblaciones de referencia del *HapMap3 project* (ver **Figura - Anexos E**) (118); lo anterior, después de aplicar los parámetros para reducir la redundancia de los SNPs. El MDS se realizó en PLINK (119) y para los análisis de ancestría se usó el algoritmo de ADMIXTURE (115), fijando el número de componentes ancestrales a k = 3.

Poblaciones de referencia: europeos [n = 112; CEU = residentes de Utah con ancestría europea], asiáticos [n = 84; CHB = asiáticos de Beijing, China] y africanos [n = 90; LWK = africanos de Webuye, Kenia].

- **Genotipificación mediante un microarreglo de genoma completo (GWAS) conocido como *"Infinium® OmniExpressExome Array".*** Esta base consiste en datos de 720815 SNPs en 415 muestras colombianas (ver **Tabla - Anexos A** y **Figura - Anexos D**); lo anterior después de aplicar los controles de calidad según el protocolo de *Anderson C, et al* (117). La base de datos consolidada y preparada para evaluar la estratificación poblacional, y estimar la ancestría global, incluyó los genotipos de 9868 SNPs comunes entre las muestras colombianas y las poblaciones de referencia de las bases públicas de *1000 Genomes* (113) y del *HGDP* (114) (ver **Figura - Anexos F**); lo anterior, después de aplicar los parámetros para reducir la redundancia de los SNPs. Tal como en el caso anterior, el MDS se realizó en PLINK (119) y los análisis de ancestría en ADMIXTURE (115), fijando el número de componentes ancestrales a k = 3.

  Poblaciones de referencia: 1000 Genomes (europeas [n = 107; IBS = población de España] y africanas [n = 108; YRI = africanos de Ibadan, Nigeria]) y *HGDP* (amerindias [n = 108; poblaciones de nativos americanos, AME= Pima, Maya, Karitiana, Surui y de Colombia]).

- **Estimación de la ancestría global a partir de la inferencia de la ancestría local (LAI) en las 415 muestras colombianas genotipadas con el microarreglo tipo GWAS.** La base de datos limpia del método anterior, se usó para realizar la inferencia de la ancestría local (LAI) en 415 muestras del estudio (ver **Tabla - Anexos A** y **Figura - Anexos D**), usando las poblaciones de referencia ya mencionadas de *1000 Genomas* (113) y del *HGDP* (114). La base de datos consolidada de las muestras colombianas con las poblaciones de referencia para LAI, incluyó los genotipos de 275284 SNPs comunes (ver **Figura - Anexos G**). Debido a que el programa usado para inferir las ancestrías locales, RFMix v1.5.4 (120), toma ventaja de los **patrones haplotípicos**, no se realizaron pasos para reducir la redundancia de los SNPs. Por último, se calculó el promedio de las ancestrías locales obtenidas para cada componente (europeo, amerindio y africano) en autosomas, con el fin de estimar la ancestría global a este nivel.

Poblaciones de referencia: *1000 Genomes* (europeas [n = 107; IBS = población de España] y africanas [n = 108; YRI = africanos de Ibadan, Nigeria]) y *HGDP* (amerindias [n = 108; poblaciones de nativos americanos, AME= Pima, Maya, Karitiana, Surui y de Colombia]).

Se realizaron correlaciones de Pearson *(r)* para cada componente ancestral en las muestras sobrelapadas, entre las tres metodologías descritas, así:

- Correlación entre las ancestrías globales estimadas con dos algoritmos o programas bioinformáticos diferentes, ADMIXTURE (115) versus RFMix (120), usando diferente número de SNPs, en 415 muestras comunes genotipadas con GWAS.
- Correlación en 85 muestras sobrelapadas, entre las ancestrías obtenidas con RFMix (120) en muestras genotipadas con GWAS versus las estimaciones obtenidas con ADMIXTURE (115) en las muestras genotipadas con CG.
- Correlación entre las ancestrías globales estimadas en ADMIXTURE (115) con dos plataformas diferentes, CG versus GWAS, usando diferente número de SNPs y poblaciones de referencia, en 85 muestras comunes.

Para evaluar las diferencias entre las proporciones de ancestrías de los casos (CCR y PA) en comparación con los controles, se usó el test no paramétrico Wilcoxon-Mann-Whitney y un test basado en 100 mil permutaciones (116). Se aplicó una **transformación logit** a las proporciones de ancestría con el fin de lograr una distribución simétrica en los análisis de regresión ajustados y se tomó como referencia la ancestría amerindia. En todos los modelos de regresión logística que incluyeron las proporciones de ancestría de 813 individuos únicos, se usó la variable *"Array"* para corregir los análisis por la plataforma usada para calcular las ancestrías globales. En los modelos de regresión multinomial de los fenotipos PA y CCR se usó el grupo *"control"* como referencia, para evaluar su asociación con los componentes ancestrales y otras variables como edad, sexo, nivel educativo, ciudad de origen, consumo de AINES e historia familiar de CCR. Se escogió el mejor modelo según el **criterio de información de Akaike (AIC)**.

1.2.3 Resultados

Análisis descriptivo de casos y controles colombianos incluidos en el estudio

Comparados con los controles, una mayor proporción de casos de PA y CCR estuvieron entre los 50 a 69 años de edad ($P < 0.01$) y la proporción de hombres fue mayor en el grupo de CCR ($P = 0.02$) (ver **Tabla 10**). Los casos de PA se caracterizaron por tener mayor nivel educativo ($P = 0.02$), mientras que los casos de CCR se caracterizaron por un menor nivel académico ($P < 0.01$). No se encontraron diferencias en cuanto a los antecedentes familiares en primer grado de CCR, ni consumo de AINES, ni región de origen, entre los grupos de estudio.

Tabla 10 Características de casos y controles incluidos

CARACTERÍSTICAS	CONTROLES		CASOS					
			PÓLIPOS ADENOMATOSOS (PA)			CÁNCER COLORRECTAL (CCR)		
	n = 506	(%)	n = 200	(%)	*P*	n = 313	(%)	*P*
Rango de edad								
30-39	103	(20.4)	9	(4.5)		18	(5.8)	
40-49	125	(24.7)	30	(15.0)		55	(17.6)	
50-59	125	(24.7)	64	(32.5)		93	(29.7)	
60-69	105	(20.8)	68	(34.0)		104	(33.2)	
70-79	48	(9.5)	29	(14.5)	**<0.01**	43	(13.7)	**<0.01**
Sexo								
Femenino	293	(57.9)	104	(52.0)		155	(49.5)	
Masculino	213	(42.1)	96	(48.0)	0.18	158	(50.5)	**0.02**
Nivel educativo								
Sin educación	8	(1.6)	5	(2.5)		19	(6.1)	
Primaria	169	(33.5)	54	(27.0)		142	(45.4)	
Secundaria	180	(35.6)	73	(36.5)		99	(31.6)	
Técnico	75	(14.9)	21	(10.5)		21	(6.7)	
Universidad o mayor	73	(14.5)	47	(23.5)	**0.02**	32	(10.2)	**<0.01**
Historia familiar de CCR								
No	328	(64.8)	119	(59.5)		212	(67.7)	
Si	178	(35.2)	81	(40.5)	0.22	101	(32.3)	0.44
Consumo de AINES								
No	394	(77.9)	159	(79.5)		252	(80.5)	
Si	112	(22.1)	41	(20.5)	0.71	61	(19.5)	0.42
Región de origen								
Andina	231	(45.7)	94	(47.0)		140	(44.7)	
Costera	275	(54.3)	106	(53.0)	0.81	173	(55.3)	0.85

Los valores de *P* corresponden a la prueba X^2 de Pearson para evaluar las diferencias entre los grupos, para las variables incluidas. Los análisis se realizaron en 1019 muestras incluidas en el estudio.

Proporciones de ancestría globales a nivel de autosomas entre casos y controles colombianos

Los gráficos MDS de las muestras genotipadas con CG y GWAS, corresponden a las **Figura 7**A y **Figura 7**C, respectivamente. En ambos casos, la mayoría de los individuos colombianos incluidos se ubicaron entre las poblaciones de referencia, europeas y asiáticas/amerindias; sin embargo, la mayor similitud fue con europeos. También se observaron algunas muestras cercanas a las poblaciones africanas de referencia. Es evidente el sobrelapamiento entre casos y controles, lo que permite concluir que los grupos a analizar en este estudio son comparables.

Las estimaciones de ancestría global por individuo obtenidas con ADMIXTURE (115) para las dos plataformas, CG y GWAS, se muestran en la **Figura 7**B y **Figura 7**D, respectivamente; adicionalmente, se muestran las correspondientes distribuciones de las proporciones de cada ancestría en las poblaciones de referencia y las de Colombia, en la **Tabla 11** y **Tabla 12**. Según las estimaciones obtenidas con la plataforma CG (ver **Tabla 11**), se observó poco mestizaje en las poblaciones CEU correspondiente a residentes de Utah con ancestría europea, y en su mayoría mormones, cuyo principal componente es el europeo (83%); mientras que las poblaciones CHB de China y LWK de Kenia se mostraron más puras, con un 93% de ancestría asiática y 94% de ancestría africana, respectivamente. En la población latina incluida como control, MEX, correspondiente a residentes de California descendientes de mexicanos, las proporciones de ancestría europea, amerindia y africana fueron 56%, 37% y 7%, respectivamente. Si bien en las muestras de Colombia a estudio, al igual que las MEX, se observa un mayor componente europeo (56%) y amerindio (27%), el componente africano es mayor en nuestras poblaciones (18%). Por otro lado, según las estimaciones obtenidas con la plataforma GWAS (ver **Tabla 12**), se observó que las muestras IBS de España y YRI de Nigeria, son prácticamente puras, con proporciones de ancestrías de 97% europea y 100% africana, respectivamente. En las poblaciones de referencia nativas americanas, AME, el componente predominante fue el amerindio con un 93%, aunque mostraron un leve grado de mestizaje con ancestros europeos del 6%. Las proporciones de ancestría europea, amerindia y africana obtenidas con esta plataforma para las muestras colombianas fueron 55%, 32% y 12%, similares a las obtenidas con la plataforma CG.

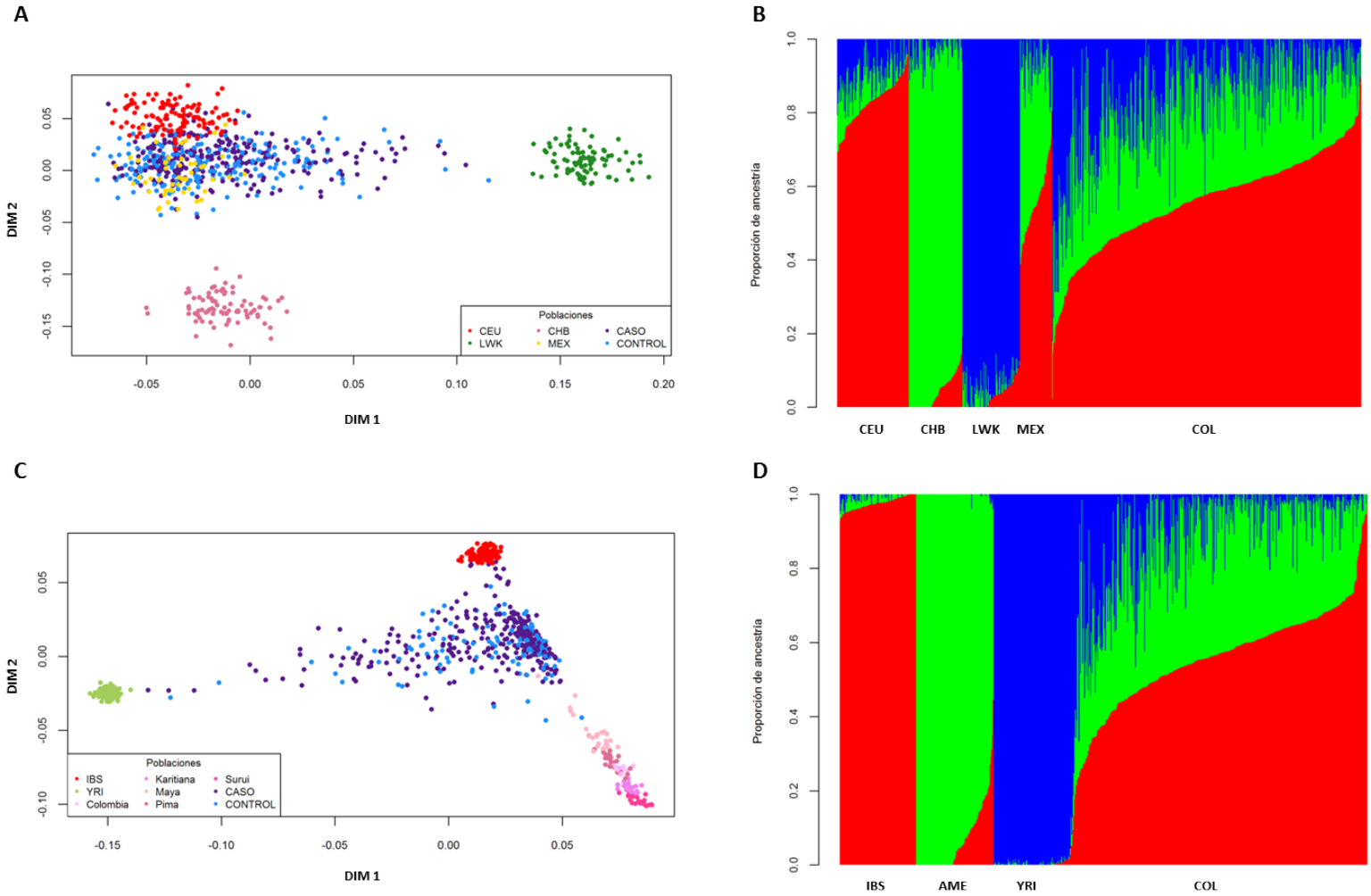


Figura 7 Análisis MDS y estimaciones de ancestría globales de muestras colombianas genotipadas con las dos plataformas. A) MDS de casos y controles genotipados con la plataforma CG, usando poblaciones de referencia del *HapMap3 project* (CEU por europeos, CHB por asiáticos/amerindios y LWK por africanos); se incluyó como control otra población mezclada, MEX = mexicanos. B) Ancestría global por individuo para las muestras genotipadas con la plataforma CG. C) MDS de casos y controles genotipados con la plataforma GWAS, usando poblaciones de

referencia de las bases de datos *1000 Genomes* (IBS por europeos y YRI por africanos) y *HGDP* (poblaciones de nativos americanos, AME: Pima, Maya, Karitiana, Surui y de Colombia). D) Ancestría global por individuo para las muestras genotipadas con la plataforma GWAS. Los análisis se realizaron con las muestras que pasaron los controles de calidad (ver **Tabla - Anexos A**) y los respectivos genotipos en autosomas (ver **Figura - Anexos E** y **Figura - Anexos F**). Figura original generada en el programa *R statistics* (76).
CEU, residentes de Utah con ancestría europea; CHB, asiáticos de Beijing, China; LWK, africanos de Webuye, Kenia; MEX, residentes de California descendientes de mexicanos; IBS, población de España; YRI, africanos de Ibadan, Nigeria; AME, poblaciones nativas americanas Pima, Maya, Karitiana, Surui y de Colombia; DIM, dimensiones; COL, muestras CASOS y CONTROL de Colombia a estudio.

Tabla 11 Distribución de las proporciones de ancestrías obtenidas con la plataforma CG en las poblaciones de referencia y de Colombia a estudio

Plataforma "CG" *	DISTRIBUCIÓN DE LAS PROPORCIONES DE ANCESTRÍAS		
	Europea	Amerindia**	Africana
Poblaciones			
CEU (n = 112)			
Mínimo	0.69	0.00	0.00
1er cuartil	0.79	0.03	0.07
Mediana	0.83	0.07	0.10
Media [95% CI]	0.83 [0.82-0.84]	0.07 [0.06-0.08]	0.10 [0.09-0.11]
3er cuartil	0.86	0.11	0.13
Máximo	1.00	0.18	0.22
CHB (n = 84)			
Mínimo	0.00	0.75	0.00
1er cuartil	0.00	0.90	0.00
Mediana	0.02	0.93	0.03
Media [95% CI]	0.04 [0.03-0.05]	0.93 [0.92-0.94]	0.03 [0.03-0.04]
3er cuartil	0.07	0.95	0.05
Máximo	0.19	1.00	0.14
LWK (n = 90)			
Mínimo	0.00	0.00	0.85
1er cuartil	0.00	0.00	0.91
Mediana	0.02	0.01	0.95
Media [95% CI]	0.03 [0.02-0.04]	0.03 [0.02-0.03]	0.94 [0.94-0.95]
3er cuartil	0.05	0.05	0.97
Máximo	0.11	0.12	1.00
MEX (n = 50)			
Mínimo	0.39	0.10	0.00
1er cuartil	0.49	0.32	0.03
Mediana	0.55	0.37	0.05
Media [95% CI]	0.56 [0.53-0.59]	0.37 [0.34-0.40]	0.07 [0.05-0.08]
3er cuartil	0.61	0.45	0.09
Máximo	0.80	0.61	0.20
COL (n = 483)			
Mínimo	0.02	0.00	0.00
1er cuartil	0.48	0.20	0.06
Mediana	0.58	0.27	0.14
Media [95% CI]	0.56 [0.55-0.57]	0.27 [0.26-0.28]	0.18 [0.16-0.19]
3er cuartil	0.64	0.34	0.26
Máximo	0.90	0.59	0.82

* Se usaron marcadores en autosomas de la base CG consolidada y reducida (ver **Figura - Anexos E**).

** Teniendo en cuenta la similitud en las frecuencias alélicas entre asiáticos y amerindios (121, 122), se usó una población de China, CHB, para inferir el componente amerindio de las muestras.

CEU, residentes de Utah con ancestría europea; CHB, asiáticos de Beijing, China; LWK, africanos de Webuye, Kenya; MEX, residentes de California descendientes de mexicanos; COL, muestras colombianas a estudio.

Tabla 12 Distribución de las proporciones de ancestrías obtenidas con la plataforma GWAS en las poblaciones de referencia y de Colombia a estudio

Plataforma "GWAS" *	DISTRIBUCIÓN DE LAS PROPORCIONES DE ANCESTRÍAS		
	Europea	Amerindia	Africana
Poblaciones			
IBS (n = 107)			
Mínimo	0.93	0.00	0.00
1er cuartil	0.96	0.01	0.00
Mediana	0.98	0.02	0.00
Media [95% CI]	0.97 [0.97-0.98]	0.02 [0.02-0.02]	0.01 [0.00-0.01]
3er cuartil	0.99	0.03	0.01
Máximo	1.00	0.05	0.05
AME (n = 108)			
Mínimo	0.00	0.51	0.00
1er cuartil	0.00	0.89	0.00
Mediana	0.02	0.98	0.00
Media [95% CI]	0.06 [0.05-0.08]	0.93 [0.92-0.95]	0.00 [0.00-0.00]
3er cuartil	0.11	1.00	0.00
Máximo	0.45	1.00	0.04
YRI (n = 108)			
Mínimo	0.00	0.00	0.96
1er cuartil	0.00	0.00	0.99
Mediana	0.00	0.00	1.00
Media [95% CI]	0.00 [0.00-0.00]	0.00 [0.00-0.00]	1.00 [0.99-1.00]
3er cuartil	0.00	0.00	1.00
Máximo	0.03	0.02	1.00
COL (n = 415)			
Mínimo	0.05	0.03	0.00
1er cuartil	0.48	0.26	0.02
Mediana	0.57	0.33	0.04
Media [95% CI]	0.55 [0.54-0.57]	0.32 [0.31-0.33]	0.12 [0.11-0.14]
3er cuartil	0.65	0.40	0.18
Máximo	0.97	0.72	0.92

* Se usaron marcadores en autosomas de la base GWAS consolidada y reducida (ver **Figura - Anexos F**).
IBS, población de España; YRI, africanos de Ibadan, Nigeria; AME, poblaciones nativas americanas Pima, Maya, Karitiana, Surui y de Colombia; COL, muestras colombianas a estudio.

De acuerdo a lo observado en la **Tabla 11** y **Tabla 12**, con respecto a la similitud en las proporciones de ancestrías en las muestras colombianas usando diferentes metodologías, encontramos que las proporciones de europeo, amerindio y africano calculadas con ADMIXTURE (115) tuvieron una alta correlación en las 85 muestras repetidas entre las dos plataformas (*r* de 0.77, 0.76 y 0.87, respectivamente; $P < 0.001$) (ver **Figura 8**C); lo anterior, aun cuando se usaron diferentes poblaciones de referencia y diferente número

de SNPs. Estas correlaciones fueron similares al comparar las estimaciones obtenidas por RFMix (120) con datos GWAS versus ADMIXTURE (115) con datos CG, para las mismas 85 muestras (ver **Figura 8**B). Por otro lado, la correlación de las ancestrías globales calculadas con los programas RFMix (120) y ADMIXTURE (115), en las 415 muestras genotipadas con la plataforma GWAS, estuvieron igual o mayor a 0.92 para todos los componentes ancestrales ($P < 0.001$) (ver **Figura 8**A).

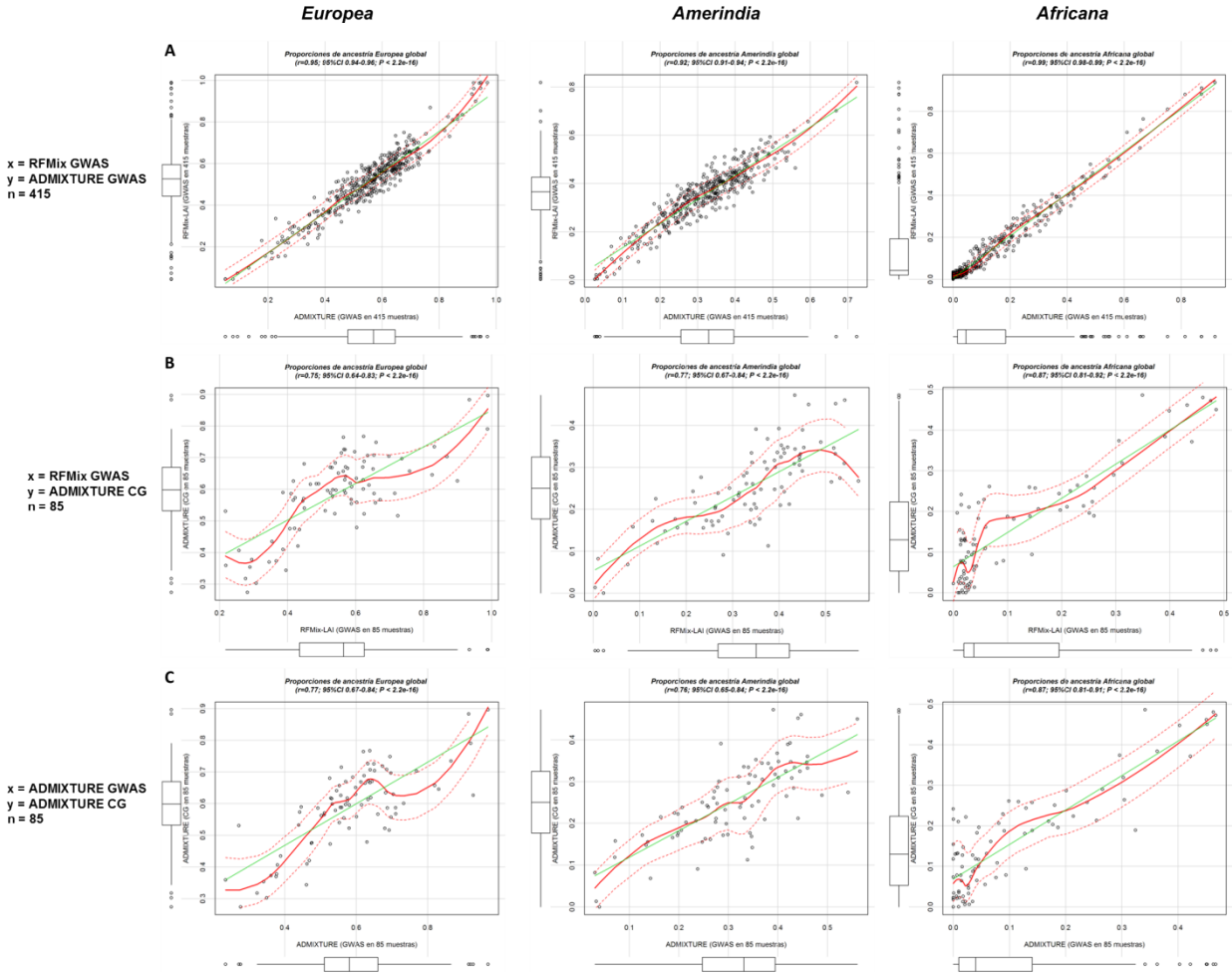


Figura 8 Correlación de Pearson *(r)* al comparar las estimaciones de ancestría globales de muestras colombianas, obtenidas mediante diferentes metodologías. A) Correlación de las ancestrías globales calculadas con los programas RFMix (120) y ADMIXTURE (115) en las 415 muestras genotipadas con la plataforma GWAS. B) Correlación de las estimaciones obtenidas con RFMix (120) en datos GWAS versus ADMIXTURE (115) en datos CG, para las 85 muestras repetidas entre las dos plataformas. C) Correlación de las estimaciones obtenidas con ADMIXTURE (115) en las 85 muestras repetidas entre las dos plataformas. Figura original generada en el programa *R statistics* (76).

Al comparar entre grupos encontramos diferencias significativas entre PA y controles, con respecto a las proporciones de ancestría europea y amerindia, obtenidas con ambas plataformas, CG (P = 0.04; $P_{100\ mil\ permutaciones}$ = 0.04 y P < 0.001; $P_{100\ mil\ permutaciones}$ < 0.01, respectivamente) (ver **Figura 9**A) y GWAS ($P \leq 0.001$; $P_{100\ mil\ permutaciones}$ < 0.01 para ambos componentes) (ver **Figura 9**B). Con respecto a la ancestría africana, se encontraron diferencias significativas entre CCR versus controles, solo en las muestras analizadas con GWAS (P = 0.07; $P_{100\ mil\ permutaciones}$ = 0.03) (ver **Figura 9**B).

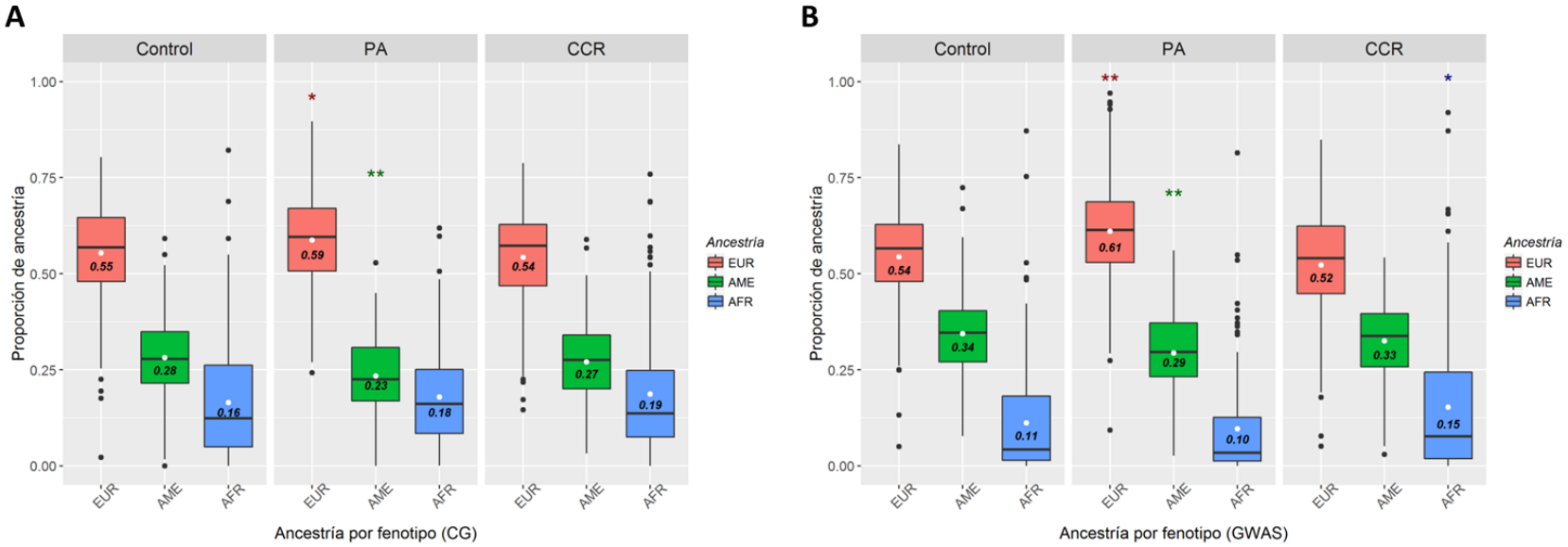


Figura 9 Diferencias en las proporciones de ancestría global, entre de PA y CCR con controles, obtenidas con las dos plataformas. Gráficos tipo cajas y bigotes que permite observar la distribución de las proporciones de cada ancestría por grupo. A) Estimaciones de ancestría estratificadas por fenotipo para las muestras genotipadas con la plataforma CG. B) Estimaciones de ancestría estratificadas por fenotipo para las muestras genotipadas con la plataforma GWAS. Los puntos blancos indican la *media* de la distribución (con sus respectivos valores anotados). Figura original generada en el programa *R statistics* (76).

** Valores de P significativos ($P_{100\ mil\ permutaciones}$ < 0.01), calculados con un test de permutaciones.

* Valores de P significativos ($P_{100\ mil\ permutaciones}$ < 0.05), calculados con un test de permutaciones.

EUR, europeo (rojo); AME, amerindio (verde); AFR, africano (azul)

Efecto de factores no genéticos en el riesgo de PA y CCR en colombianos con diferentes proporciones de ancestría

Se realizaron modelos de regresión logística multinomial ajustados en 813 individuos únicos de los cuales obtuvimos las estimaciones de ancestría global usando las dos plataformas (ver

Tabla 13). Los análisis buscan explicar el riesgo de PA y CCR por diferentes variables como sexo, edad, nivel educativo, ciudad de origen, historia familiar de CCR y consumo de AINES, en colombianos con diferentes proporciones de ancestría, con el fin de obtener evidencia sobre el efecto de estas variables en tumores colorrectales en las poblaciones colombianas.

El mejor modelo según el parámetro AIC, incluyó las proporciones de ancestrías europea y africana junto con *"Array",* sexo, edad, nivel educativo y consumo de AINES (modelo 13; ver

Tabla 13). Se evidenció que la ancestría europea se asoció al riesgo de PA (OR 2.03; 95%CI 1.40-2.93; $P < 0.01$), mientras que la ancestría africana se asoció tanto al riesgo de PA (OR 1.12; 95%CI 1.03-1.21; $P = 0.01$) como de CCR (OR 1.10; 95%CI 1.03-1.17; $P = 0.01$); además de la asociación con las ancestrías, otros factores como la edad, el nivel educativo y el consumo de AINES, explicaron parte del riesgo de PA y CCR (ver **Tabla 14**).

Particularmente, la asociación de la edad a un mayor riesgo de PA fue significativa (OR 1.02; 95%CI 1.00-1.05; $P = 0.02$), pero solo se vio una tendencia con respecto al riesgo de CCR (OR 1.01; 95%CI 0.99-1.03). Por otro lado, el consumo de AINES de al menos una vez por semana por mínimo seis meses seguidos, en los últimos dos años al momento del diagnóstico, se observó como factor protector para CCR (OR 0.66; 95%CI 0.44-0.98; $P = 0.04$), pero no para PA (ver **Tabla 14**).

Tabla 13 Modelos de regresión logística multinomial para evaluar el efecto de variables de riesgo no genéticos y la ancestría en el desarrollo de PA y CCR en los colombianos

Modelos de regresión logística multinomial	Df	AIC
Modelo 1: Fenotipo ~ Sexo	4	1727.500
Modelo 2: Fenotipo ~ Edad	4	1721.783
Modelo 3: Fenotipo ~ Sexo + Edad	6	1722.519
Modelo 4: Fenotipo ~ Sexo + Edad + Edu	14	1699.342
Modelo 5: Fenotipo ~ logit(EUR) + logit(AFR) + Array	8	1645.977
Modelo 6: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo	10	1647.641
Modelo 7: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad	12	1647.551
Modelo 8: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad + Edu	20	1631.367
Modelo 9: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad + Edu + Ciu	30	1632.505
Modelo 10: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad + Edu + Ciu + AINES	32	1632.403
Modelo 11: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad + Edu + Ciu + AINES	34	1635.075
Modelo 12: Fenotipo ~ logit(EUR) + logit(AFR) + Array+ Sexo + Edad + Edu + Ant_F	22	1632.816
Modelo 13: Fenotipo ~ logit(EUR) + logit(AFR) + Array + Sexo + Edad + Edu + AINES	**22**	**1630.806**

Todos los modelos se corrieron en *R statistics*. El Modelo 13 fue el mejor y menos complejo *(Df 22; AIC 1630.806)*; los resultados se describen en la **Tabla 14**.
Fenotipo (Control, PA o CCR); Edu, nivel educativo; logit(EUR), ancestría europea transformada; logit(AFR), ancestría africana transformada; *Array*, genes candidatos (CG) o genoma completo (GWAS); Ciu, ciudad de origen; AINES, consumo de antiinflamatorios no esteroideos; Ant_F, historia familiar de CCR; Df, grados de libertad; AIC, criterio de información de Akaike.

Con respecto al nivel educativo, un indicador indirecto y proporcional del nivel socioeconómico, se observa que a mayor educación disminuye el riesgo de CCR (primaria OR 0.35, secundaria OR 0.32 y técnico OR 0.21; $P < 0.05$); estos resultados son similares en el grupo con estudios universitarios o mayores, aunque con valores de significancia marginal (ver **Tabla 14**). Con el fin de estudiar mejor el papel del nivel socioeconómico en el riesgo de tumores colorrectales en los colombianos, se realizó una regresión logística multinomial del nivel educativo explicado por las variaciones en las proporciones de ancestrías; al respecto se obtuvo que los individuos con mayor educación se caracterizan por tener mayor ancestría europea (OR's 2.08-4.90; $P \leq 0.01$), mientras que no se encontraron diferencias significativas con la ancestría africana (ver **Tabla 15**).

Tabla 14 Asociación de la ancestría global y otras variables con el riesgo de PA y CCR

CARACTERÍSTICAS	PÓLIPOS ADENOMATOSOS (PA)			CÁNCER COLORRECTAL (CCR)		
	OR	[95%CI]	*P*	OR	[95%CI]	*P*
Proporción de ancestría						
Africana*	1.12	[1.03-1.21]	**0.01**	1.10	[1.03-1.17]	**0.01**
Europea*	2.03	[1.40-2.93]	**<0.01**	1.05	[0.78-1.41]	0.77
Sexo						
Femenino	1	ref	ref	1	ref	ref
Masculino	1.03	[0.69-1.53]	0.90	1.20	[0.87-1.67]	0.26
Edad	1.02	[1.00-1.05]	**0.02**	1.01	[0.99-1.03]	0.20
Nivel educativo						
Sin educación	1	ref	ref	1	ref	ref
Primaria	0.34	[0.10-1.18]	0.09	0.35	[0.14-0.90]	**0.03**
Secundaria	0.58	[0.17-2.04]	0.40	0.32	[0.12-0.83]	**0.02**
Técnico	0.55	[0.14-2.16]	0.39	0.21	[0.07-0.63]	**<0.01**
Universidad o mayor	1.23	[0.33-4.59]	0.76	0.39	[0.14-1.09]	0.07
Consumo de AINES						
No	1	ref	ref	1	ref	ref
Sí	0.73	[0.45-1.17]	0.19	0.66	[0.44-0.98]	**0.04**

Los valores de *P* corresponden a la regresión logística multinomial para evaluar el efecto del componente ancestral en el riesgo de tumores colorrectales en 813 muestras de Colombia genotipadas con dos plataformas, CG o GWAS. El modelo está ajustado por sexo, edad, nivel educativo y consumo de AINES (mejor modelo; ver

Tabla 13).

*Proporciones de ancestrías bajo transformación logit.

Tabla 15 Proporciones de ancestría en los diferentes niveles educativos de los colombianos

ANCESTRÍAS	NIVEL EDUCATIVO (Indicador del estrato socio-económico)				
	Ninguno	Primaria	Secundaria	Técnico	Universitario o mayor
Africana*	ref	0.89 [0.73-1.10] *P* = 0.28	0.99 [0.80-1.21] *P* = 0.89	0.94 [0.75-1.17] *P* = 0.58	0.98 [0.79-1.22] *P* = 0.87
Europea*	ref	2.08 [1.17-3.72] ***P* = 0.01**	2.61 [1.44-4.71] ***P* < 0.01**	2.39 [1.20-4.76] ***P* = 0.01**	4.90 [2.52-9.50] ***P* < 0.01**

Los valores de *P* corresponden a la regresión logística multinomial para evaluar diferencias en las proporciones de ancestrías según el nivel educativo en 813 muestras de Colombia genotipadas con las plataformas, CG o GWAS.

*Proporciones de ancestrías bajo transformación logit.

1.3 Discusión

Actualmente la evaluación del componente étnico, medido objetivamente como proporciones de ancestría, es un aspecto muy importante en el diseño adecuado de estudios de asociación poblacionales dirigidos a la identificación de marcadores genéticos de susceptibilidad en enfermedades complejas, como el CCR; especialmente, si la población se caracteriza por un alto grado de mestizaje (70).

Dentro de los objetivos de este primer capítulo, estuvieron incluidos medir las proporciones de ancestría y evaluar la presencia de sesgos de género en el proceso de mestizaje en las poblaciones colombianas analizadas en este estudio, con el fin de obtener evidencia científica de las diferencias descritas en nuestra historia desde la colonización. Lo anterior, con el propósito de evaluar el efecto de la estructura genética de los colombianos en el riesgo de neoplasias colorrectales, en modelos de regresión logística corregidos por otros factores de riesgo no genéticos.

Para evaluar lo anterior, se incluyeron ciudades representantes de varios departamentos de Colombia ubicados en la región andina y costera, que se caracterizan por una composición étnica y proporciones de ancestrías europea, amerindia y africana variables, según los antecedentes históricos y estudios realizados en marcadores genéticos, respectivamente. Estas ciudades fueron Bogotá DC y Bucaramanga de la región andina, y Cali, Barranquilla, Cartagena y Santa Marta de la región costera del país.

Dentro de estas poblaciones, Bogotá DC y Cali en Valle del Cauca pueden ser considerados como de alto riesgo para CCR, pues se observó un aumento del 22% en el riesgo de muerte con respecto al resto de la población colombiana (2). Igualmente, Bucaramanga en Santander podría ser considerada de alto riesgo de acuerdo a los análisis de tendencia de la mortalidad por CCR, que reportan un incremento anual promedio del 2.3% y 2.4% en hombres y mujeres, respectivamente; siendo mayor que lo reportado en la población general de Colombia (2.2% y 1.9% en hombres y mujeres, respectivamente) (2). A pesar de que valores de REM por debajo de 100 son difíciles de interpretar, podemos decir que según las TAE en incidencia y mortalidad por CCR en el país, Cartagena en Bolívar y Santa Marta en Magdalena, podrían considerarse de bajo

riesgo para CCR, mientras que Barranquilla en el Atlántico podría clasificarse como de riesgo moderado (111).

Analizando una submuestra de 131 colombianos *"controles"* (con genotipos a nivel de genoma completo, ~ 9800 SNPs) de seis ciudades del país, se obtuvieron estimaciones de ancestrías europea, amerindia y africana de 54%, 34% y 11% en autosomas y 31%, 54% y 14% en cromosoma X, respectivamente. En general, estos resultados están en línea con lo reportado en otros estudios realizados en poblaciones colombianas del Cauca (4), región Andina Oriental (Boyacá, Santander, Norte de Santander y Huila) (6), región de la Orinoquía (Arauca, Casanare y Meta) (6), Magdalena, Caldas, Quindío y Santander (10), y otras no especificadas (85), en los cuales también se ha observado una mayor proporción del componente europeo en autosomas junto con un mayor componente amerindio en el cromosoma X. Otras publicaciones en población de Antioquia, han reportado de manera consistente que la ancestría europea es predominante tanto en autosomas como en el cromosoma X (3, 6, 11); estos resultados no son comparables con los resultados de esta tesis doctoral, pues no se incluyó ninguna población de este departamento.

En los análisis estratificados por ciudad y región, se observó la misma tendencia aunque estas diferencias solo fueron significativas para los componentes europeo y amerindio en Bogotá DC, Bucaramanga, Barranquilla y Cartagena; también, en Cali se encontraron diferencias significativas en el componente ancestral europeo en autosomas versus cromosoma X, mas no en el amerindio. Con respecto a las diferencias en la ancestría africana en autosomas versus cromosoma X, solo fue significativa en las ciudades de la región andina. La principal razón por la que posiblemente no se obtuvieron valores de *P* significativos de manera consistente en las muestras de la Costa Pacífica y Caribe, al comparar las ancestrías en autosomas versus cromosoma X, posiblemente está relacionada con el tamaño de la muestra cuando se realizaron los análisis estratificados por ciudad.

En las ciudades de la región costera llama la atención que de la población efectiva europea, más del 95% eran hombres en comparación con un 78% en la región andina; es decir, en las ciudades de la Costa se observa una mayor asimetría en la contribución de género en europeos que en ciudades de la región andina. Lo anterior, está en acuerdo

con la historia del país desde la colonización, que describe la llegada principalmente de hombres europeos que primero fundaron Santa Marta y Cartagena a inicios del siglo XVI, y que después se distribuyeron a lo largo del río Magdalena fundando Bogotá pocos años después (74). Una vez hubo asentamiento de los hombres europeos, llegaron mujeres y familias del viejo continente que se ubicaron principalmente en zonas urbanas del interior del país, razón por la cual se evidencia una mayor participación de mujeres europeas en el proceso de mestizaje en la región andina (22%) versus la región costera (~2% a 5%).

De la población efectiva amerindia el porcentaje de mujeres varió menos entre las ciudades, siendo menor en Barranquilla y Cartagena con un 73%, seguido de las ciudades de la región andina con ~77% y, Santa Marta y Cali con un ~80%. Teniendo en cuenta que entre 9.000 a 1.500 años a.C se llevó a cabo la dispersión y asentamiento de nativos americanos en América del Sur, formando diferentes comunidades que se establecieron a lo largo del territorio nacional desde antes de la llegada de los europeos (73), no es de extrañar que se encontrara una contribución relativamente homogénea de mujeres amerindias en el proceso de mestizaje en las dos regiones incluidas en este estudio.

En las ciudades de la región andina se encontró una leve mayor contribución de mujeres (~58%) que de hombres africanos (~42%) en el proceso de mestizaje, lo que está en acuerdo con la trata de mujeres de origen africano para el servicio doméstico en familias europeas del interior del país, mientras que los hombres fueron principalmente dispuestos a trabajar en zonas mineras no urbanas de los Andes y del Pacífico, aunque también realizaron tareas pesadas en las haciendas pertenecientes a europeos (75). Si bien en las ciudades de la región costera en general se observó una mayor contribución de la raza negra en el proceso de mestizaje (21%), en comparación con la región andina (3%) según resultados a nivel de genoma completo, las contribuciones de género fueron variables en las ciudades de la Costa. Por ejemplo, en Barranquilla y Cartagena la población efectiva africana fue en su mayoría hombres (con un 78% y 67%, respectivamente), mientras que en Cali y Santa Marta la contribución africana se dio principalmente por mujeres (con un 81% y 89%, respectivamente).

Nuestros resultados con respecto a la estructura genética de las poblaciones colombianas incluidas en este estudio, son coherentes con la historia de nuestro país

puesto que muestran que hubo asimetría en la proporción de hombres y mujeres de cada población continental en el modelamiento de las poblaciones actuales colombianas; lo anterior, aun cuando en los análisis de sesgos de género se asumió una población mezclada con igual número de hombres que de mujeres, que surgió de un evento de mestizaje en un momento en el tiempo, sin tener en cuenta el número de generaciones desde el evento, el cual está estimado en $\lambda = 9{,}2 \pm 0.9$ (8).

Por otro lado, vale la pena mencionar que las variaciones observadas entre los estudios, incluido este trabajo de tesis, con respecto a las proporciones ancestrales estimadas, incluso para una misma región, pueden deberse a: i) el número de marcadores usado para hacer las estimaciones; ii) las poblaciones de referencia incluidas; iii) el programa bioinformático implementado; y iv) los criterios de selección de la muestra a estudiar.

En relación con otros estudios publicados, es de resaltar en este trabajo de tesis que las estimaciones de ancestría con las cuales posteriormente se analizaron los posibles sesgos de género en el proceso de mestizaje en colombianos, se obtuvieron usando más de 9 mil SNPs, comparable solo con el estudio de *Rishishwar L et al.,* en el cual usaron datos de genoma completo para estimar las ancestrías en población Antioqueña (9). Otro punto importante, es que las estimaciones se realizaron en simultáneo para una muestra de 415 colombianos, aun cuando después se seleccionaron solo los *"controles"* para evaluar la asimetría de género en el proceso de mestizaje. Vale la pena aclarar que la decisión de solo usar *"controles"* para estos análisis, se debió a que teníamos evidencia previa en nuestro grupo acerca de la asociación de las ancestrías europea y africana con el riesgo de tumores colorrectales (121) y porque los *"controles"* representan mejor la población general colombiana. Finalmente, se destaca que se usaron poblaciones de referencia de bases de datos reconocidas, con representación de poblaciones puras de España y África, junto con poblaciones amerindias (incluyendo nativos americanos de Colombia), las cuales consideramos óptimas para hacer las estimaciones de ancestrías en poblaciones Latinas, como Colombia.

Se puede concluir de esta primera parte del estudio, que existen diferencias en las proporciones de ancestría europea, amerindia y africana en las poblaciones colombianas seleccionadas y que estas diferencias son el resultado del patrón de mestizaje que se llevó a cabo en estas regiones, el cual estuvo condicionado por el modelo migratorio de

las poblaciones continentales que se asentaron a lo largo del territorio nacional y el rol que éstos ancestros asumieron durante el periodo de la colonización en Colombia. Ahora bien, con respecto al estudio del efecto de estas diferencias en las proporciones de ancestrías sobre la variación en el riesgo de CCR en los colombianos, también se obtuvieron resultados muy interesantes los cuales se discutirán a continuación, junto con los resultados con respecto al efecto de otros factores de riesgo.

En los análisis descriptivos de la muestra total de 1019 colombianos, incluidos en este estudio en un periodo de cuatro años desde el 2008 al 2011, se encontró que la edad avanzada es un factor de riesgo para PA y CCR puesto que ≥ 78% de éstos eran mayores de 50 años, en comparación con un 55% de los controles, lo cual está en línea con lo reportado en la literatura (123). Más aún, esta asociación se mantuvo en los análisis de regresión logística multinomial ajustados por otras variables en una submuestra de 813 colombianos, aunque el tamaño del efecto fue bajo.

Igualmente, se han reportado diferencias en el sexo con respecto al riesgo de CCR (123). Según *GLOBOCAN 2012*, en USA la tasa de incidencia ajustada por edad en hombres es mayor que en mujeres (28.5 versus 22.0) y en Colombia el comportamiento es similar aunque en menor proporción (13.4 versus 12.5) (1). Al respecto, en este estudio se observó un mayor número de hombres en el grupo de CCR comparado con controles (51% vs 42%), mientras que en mujeres se vio lo contrario (50% vs 58%); sin embargo, estas diferencias no se mantuvieron en los análisis de regresión multinomial ajustados.

La **heredabilidad** en CCR está estimada en el ~35% de los casos según estudios en gemelos (124), aunque solo el 5% del total está explicado por síndromes hereditarios (125). Entonces, se sabe que la historia familiar es un factor de riesgo importante en tumores colorrectales, especialmente si existen casos de cáncer en edades muy tempranas (123). A pesar de lo anterior, en este estudio no se encontraron diferencias significativas en la historia familiar de CCR entre casos y controles; esto podría tener relación con que el estudio se enfocó en los casos de CCR esporádicos, al no incluir todos los posibles cánceres hereditarios en menores de 30 años.

Si bien en los análisis descriptivos no se encontraron diferencias significativas entre el riesgo de PA y CCR con el consumo de AINES, si se evidenció su papel protector en el

riesgo de CCR en los análisis de regresión logística multinomial ajustados, que mostraron una disminución del riesgo hasta en un 34% con el consumo de AINES de al menos una vez por semana por mínimo seis meses seguidos, en los últimos dos años al momento del diagnóstico. Lo anterior está en acuerdo con publicaciones recientes que reportan una disminución de hasta un 43% en el riesgo de CCR en personas con un consumo continuo y prolongado de AINES con **inhibición selectiva** sobre la enzima COX-2 (126); más aún, se ha reportado que el uso de inhibidores de COX-2 suprimió el desarrollo de poliposis intestinal en ratones con inactivación del gen *APC* (127) y redujo los niveles de PGE2 con la consecuente inhibición del crecimiento de líneas celulares de CCR (128). Entonces, este estudio y otros soportan el rol que juega la inflamación en el proceso de tumorogénesis (129).

Un aspecto importante a resaltar, es que no se encontraron diferencias significativas entre casos y controles con respecto a la región de origen, andina y costera, es decir que no existieron sesgos en la recolección de los participantes del estudio. Lo anterior, está soportado también en los análisis de MDS realizados en los dos set de muestras colombianas, CG y GWAS, en los cuales es posible evidenciar el sobrelapamiento entre casos y controles del país, indicando que son grupos comparables.

Existe evidencia sobre el papel de la raza y la etnia en el riesgo de diferentes tipos de cáncer, incluido el CCR, la cual se expuso en el marco teórico de este primer capítulo. Teniendo en cuenta que somos una mezcla de tres diferentes razas, esta característica se midió en los colombianos del estudio de manera objetiva mediante el análisis de las proporciones de las ancestrías europea, amerindia y africana a nivel de autosomas en 813 individuos únicos del estudio.

Los análisis de ancestría se realizaron usando diferentes aproximaciones. Con la plataforma CG, se analizaron 483 muestras colombianas con 473 SNPs en el programa ADMIXTURE (115), usando 336 muestras de poblaciones de referencia. La estimación en la ancestría asiática obtenida para la población CHB fue del 93%, similar al 96% reportado por *Pardo-Seco J, et al.,* mientras que la estimación de la ancestría Europea en la población CEU fue mucho menor a la reportada por *Pardo-Seco J, et al.,* (83% versus 93%); lo anterior, aun cuando usaron un panel de tamaño similar con ~ 400 SNPs (130). Como era de esperarse, cuando usaron alrededor de 10000 SNPs, estas estimaciones

fueron mayores al 99% (130) y sus resultados están correlacionados con otro estudio en el cual usaron más de 1 millón de SNPs (131). Por otro lado, con la plataforma CG se obtuvo un 94% de ancestría africana en la población LWK, el cual es cercano al 98% reportado por *Williams RC, et al.,* usando 1300 SNPs (132). Con respecto a la población Latina incluida como control, MEX, se obtuvieron estimaciones del 56%, 37% y 7% para las ancestrías europea, amerindia y africana, respectivamente; estos resultados son muy similares a los reportados por *Ma Y, et al.,* de 54%, 36% y 10%, respectivamente, usando más de 1 millón de SNPs (131) y están en línea con los reportados por *Johnson NA, et al.,* los cuales muestran una mayor proporción de ancestría europea (49%) que amerindia (45%), junto con un bajo componente africano (5%), usando > 400 mil SNPs (133). Con la plataforma GWAS, se analizaron 415 muestras colombianas con 9663 SNPs en el programa ADMIXTURE, usando 323 muestras de poblaciones de referencia. Las estimaciones obtenidas para las poblaciones de referencia IBS y YRI, fueron de 97% europea y 100% africana, tal como se espera para estas poblaciones puras y como se ha reportado en otros estudios (134). Por otro lado, el porcentaje de ancestría amerindia obtenida para las poblaciones nativas americanas de México (Pima y Maya), Colombia (Piapoco y Curripaco) y Brasil (Karitiana y Surui), seleccionadas del HGDP fue de 95% con un 6% de ancestría europea, lo que es esperado para estas poblaciones según la literatura (135).

Solo comparando los resultados obtenidos para las poblaciones de referencia usadas con cada plataforma, podemos decir que los análisis realizados con la plataforma GWAS, son más precisos que los obtenidos con la plataforma CG. Lo anterior se debe a que usamos 20 veces más SNPs en los análisis con GWAS versus CG y también a que el uso de poblaciones nativas americanas para estimar la ancestría amerindia, es más preciso que usar la población CHB (136). Cabe resaltar que la decisión de usar la población CHB del *HapMap3 project* (118) en vez de las nativas americanas del *HGDP (114)* con la plataforma CG, se debió a que la baja densidad de SNPs en esta plataforma no permitió obtener suficientes marcadores sobrelapados entre las bases de datos. Sin embargo, es de resaltar que a pesar de estas diferencias se obtuvieron correlaciones significativas de ≥ 76% entre las estimaciones de ancestrías generadas con ADMIXTURE (115) para los 85 individuos repetidos o comunes entre las dos plataformas. Más aun, las correlaciones entre las estimaciones de ancestrías obtenidas con RFMix (120) versus ADMIXTURE (115), para las 415 muestras genotipadas con GWAS fueron ≥ 92%; todo lo

anterior es indicador de la precisión en las estimaciones de ancestría obtenidas en este estudio, especialmente con la plataforma GWAS.

Con respecto a las diferencias en las proporciones de ancestría entre los grupos de estudio, AP y CCR versus controles, encontramos una mayor ancestría europea junto con una menor ancestría amerindia en el grupo de PA comparado con los controles, en los dos sets de muestras (CG y GWAS); estos resultados fueron consistentes según los análisis de regresión logística multinomial en los cuales se evidenció la asociación de la ancestría europea en el riesgo de PA. Por otro lado, para la ancestría africana se encontraron diferencias significativas entre el grupo de CCR versus controles, solo en el set de muestras genotipadas con GWAS; sin embargo, en los análisis de regresión multinomial se encontró que efectivamente la ancestría africana estaba asociada no solo al riesgo CCR sino también de PA. Estos resultados en una muestra de 813 colombianos son consistentes con los resultados preliminares publicados por nuestro grupo en un subset de muestras (121). Más aún, las asociaciones encontradas se mantuvieron en los análisis ajustados por otras variables como sexo, edad, nivel educativo y consumo de AINES.

Una variable importante a tener en cuenta en los estudios de asociación para evaluar el efecto de la raza, etnia o ancestría en el riesgo de enfermedad, es el estrato socioeconómico; lo anterior, debido a que éste puede ser un indicador de la influencia de barreras, como el acceso a la atención médica oportuna y la falta de educación, en el control de enfermedades como el CCR a nivel de salud pública. En este estudio no se recogió la información sobre el estrato socioeconómico; sin embargo, usamos la variable de nivel educativo como un indicador de éste. Al respecto, en los análisis descriptivos se encontraron diferencias significativas entre grupos; el 71% de los casos con PA realizaron estudios de secundaria o mayor versus un 65% de los controles, mientras que este porcentaje fue del 49% en el grupo de CCR. Si bien en los análisis de regresión logística igualmente se evidenció que los individuos con mayor educación presentaron menos CCR al compararlos con aquellos sin educación, no se observó ninguna asociación entre el nivel educativo y el riesgo de PA.

Al evaluar mejor el papel del nivel educativo, encontramos que la ancestría europea tuvo una asociación positiva con la educación de los colombianos, más aún, esta asociación

se vio más fuerte en aquellos con estudios universitarios o mayor; mientras que no se observaron diferencias con respecto a la ancestría africana. Estos resultados son muy importantes en el contexto de nuestra población, pues muestra que independientemente del nivel educativo (estrato socioeconómico o acceso a la salud), existe un efecto importante de la ancestría en el riesgo de tumores colorrectales en los colombianos.

Llama la atención de estos resultados que la ancestría africana se encontró asociada tanto a CCR como a lesiones preneoplásicas colorrectales, lo cual puede indicar un rol importante de este componente en la progresión adenoma-adenocarcinoma. Por otro lado, solo se encontró una asociación de la ancestría europea con PA y no con CCR; esto puede ser posible ya que no todos los adenomas progresan a cáncer. Otra razón, es que el efecto de este componente en CCR sea más bajo y por lo tanto el tamaño de la muestra de este estudio no haya sido suficiente para detectar dicho efecto.

Los resultados con respecto a la asociación entre la ancestría africana y el riesgo de tumores colorrectales son de gran importancia, especialmente, teniendo en cuenta que la población con mayor REM para CCR fue Cali en Valle del Cauca de acuerdo al *Atlas de mortalidad por cáncer en Colombia 2000-2006* (2), y que según el *DANE* en esta ciudad se concentra el mayor porcentaje de afrocolombianos del país (82).

2. Variantes genéticas comunes asociadas al riesgo de tumores colorrectales en los colombianos

Aspectos relacionados con el estudio de la heredabilidad en CCR:

Se ha alcanzado un gran éxito en la identificación de genes de alta penetrancia involucrados en el riesgo de CCR hereditario (ver **Tabla 1**) mediante estudios de mapeo por ligamiento basados en familias *"Análisis de ligamiento en familias"* (ver **Figura 1**), los cuales han permitido encontrar la variante genética rara o mutación que se está segregando con la enfermedad dentro de una familia en la cual es posible establecer un patrón de herencia mendeliana (137-139). Si bien los estudios en familias tienen un tamaño de muestra pequeño y las variantes a detectar son poco frecuentes, aun así es posible identificar la variante causal debido al gran tamaño del efecto de éstas en el riesgo de enfermedad. A pesar de los esfuerzos en encontrar los genes implicados en el riesgo de CCR hereditario, estos solo explican el 5% de todos casos (125).

Entonces, conociendo que la mayoría de los CCR no clasifican como de tipo hereditario, pero que muchos de éstos se consideran que tienen una base genética heredable, se han implementado los estudios de asociación poblacionales, como son los GWAS, con el fin de encontrar más variantes genéticas implicadas en el riesgo de CCR (34-56). Debido al carácter poligénico y multifactorial del CCR, el principal reto de estos estudios de asociación es contar con un gran número de individuos casos y controles que permitan tener el suficiente poder estadístico para detectar las variantes asociadas al riesgo, aun cuando el tamaño del efecto sea muy pequeño, disminuyendo al máximo las asunciones debidas al azar (ver **Figura 1**) (16, 140-142). Los estudios GWAS han permitido identificar una gran cantidad de variantes, las cuales individualmente confieren un incremento

relativo del riesgo muy bajo, 1.1-1.5 veces (16); sin embargo, éstas solo explican una parte de la proporción de la heredabilidad del CCR que está estimada en un ~35% (124).

Algunas de las razones relacionadas con la ***heredabilidad desconocida*** en el estudio de las enfermedades complejas son:

i) **Una disminución en la aptitud reproductiva en determinados fenotipos asociados a alelos o genotipos específicos en un locus**, por ejemplo, variantes genéticas asociadas a desórdenes del espectro autista pueden perderse o disminuir en frecuencia en la siguiente generación, debido a la baja capacidad o aptitud reproductiva de los individuos autistas (16);

ii) **Que la selección natural puede estar favoreciendo la permanencia de variantes de bajo efecto y la eliminación de variantes de mayor efecto en las poblaciones**, por lo tanto, otro factor importante es el tamaño de la muestra de los estudios de asociación. Si la patología a estudiar es relativamente frecuente y se logran incluir suficientes casos y controles, la posibilidad de encontrar asociaciones, aún con un tamaño del efecto bajo, son mayores (16);

iii) **Que por el carácter poligénico en este tipo de enfermedades, múltiples alelos están influenciando simultáneamente el riesgo de la enfermedad**, junto con otros factores, lo cual no siempre se evalúa en los estudios de asociación afectando así la capacidad de éstos para encontrar variantes genéticas de riesgo (16);

iv) **Que el uso de plataformas que incluyen solo variantes comunes tipo SNPs, con efecto moderado o bajo, son insuficientes** en la detección de variantes funcionales y raras, que pueden tener un mayor efecto en el riesgo de la enfermedad. La implementación de técnicas de secuenciamiento de última generación de forma generalizada, contribuiría a la detección de variantes de baja frecuencia que no han sido contempladas en los estudios de asociación tradicionales de genoma completo (16);

v) **Que la mayoría de los estudios genéticos de asociación, especialmente aquellos a gran escala, se han realizado principalmente en poblaciones europeas** relativamente puras (16), mientras que las poblaciones con mezclas recientes del nuevo continente, que tienen mayor variación genética, han sido poco representadas en estos estudios (143). Precisamente, en la actualidad

existe un gran interés de la comunidad científica no solo en replicar en poblaciones Latinas aquellas variantes genéticas ya identificadas como asociadas, sino también en encontrar nuevas variantes genéticas que contribuyan a explicar la heredabilidad en enfermedades como el cáncer (91, 143, 144).

vi) **Que las variaciones genéticas con un MAF entre 0.5% y 5%, difícilmente pueden ser detectadas como asociadas a una enfermedad pues el tamaño del efecto suele ser moderado o bajo**, es decir, el tamaño del efecto no es lo suficientemente grande como para ser detectado en estudios de ligamiento en familias y tampoco son lo suficientemente frecuentes en la población como para ser detectadas en estudios de asociación de genoma completo, luego la existencia de este tipo de variantes asociadas a enfermedad junto con la ausencia de métodos para su detección, contribuyen también a la falta de conocimiento con respecto a la heredabilidad de enfermedades complejas (16).

Generalidades de los estudios genéticos de asociación a nivel poblacional en el estudio de enfermedades complejas:

Habiendo encontrado en su gran mayoría las mutaciones en genes de alta penetrancia implicadas en el desarrollo de síndromes mendelianos en cáncer, alternativas como los estudios genéticos de asociación a nivel poblacional resultan una gran opción para seguir avanzando en dilucidar la heredabilidad desconocida en cáncer; lo anterior, con el fin de contribuir en la identificación de factores de susceptibilidad genética que expliquen la variación o heterogeneidad en el riesgo de cáncer en los individuos, pues el entendimiento de estas diferencias contribuiría a mejorías en la prevención, diagnóstico y tratamiento del cáncer en general (16).

Los estudios genéticos de asociación a nivel poblacional de casos y controles buscan comparar las frecuencias alélicas y genotípicas de marcadores genéticos, usualmente SNPs, entre individuos enfermos y sanos de una población, con el fin de determinar asociaciones estadísticamente significativas entre la enfermedad y el marcador genético

(145). Existen en general dos tipos de aproximaciones para los estudios genéticos de asociación poblacionales, la de genes candidatos (CG) y la de genoma completo (GWAS).

Los estudios CG se caracterizan por el uso de paneles que incluyen varios SNPs en genes que han sido previamente relacionados con la enfermedad de interés, por ejemplo con cáncer, y que pueden ser **variantes funcionales** o marcadores en **desequilibrio de ligamiento (LD)** con las variantes funcionales, es decir **"tagSNPs"** (146). Una limitante de este tipo de estudios es que muy pocos SNPs se distribuirán bajo la hipótesis nula de *no asociación* disminuyendo la capacidad de detectar asociaciones azarosas o la influencia de **factores confusores** en los análisis (117). Más aún, al incluir un número limitado de SNPs no es posible seguir pasos muy estrictos en el control de calidad, por ejemplo: i) para la detección de SNPs con **falla en el genotipado** o con **tasas de heterocigocidad** inadecuadas; ii) o para la identificación de individuos emparentados; iii) o para evaluar adecuadamente la presencia de **subestructura poblacional**; iv) o para obtener estimaciones precisas de ancestría que permitan posteriormente corregir los análisis de asociación de la influencia de estos potenciales confusores (117). Adicionalmente, factores como el uso de tamaños de muestra pequeños o moderados, el pobre cubrimiento de la variación genética en estos paneles y la inclusión de solo algunos genes candidatos, han igualmente limitado la obtención de resultados reproducibles.

De otra parte, los estudios GWAS se realizan típicamente sin una hipótesis previa con respecto al efecto del locus en el riesgo de enfermedad, razón por la cual se basan en el uso de paneles que incluyen múltiples SNPs, alrededor de un millón, representativos de todo el genoma humano, tanto de las regiones codificantes como no codificantes (146). Debido al gran número de marcadores incluidos, este tipo de estudios requieren de un proceso de control de calidad muy estricto con el fin de eliminar los SNPs o los individuos que fallaron en el genotipado; además, es preciso hacer correcciones por múltiples test con el fin de evitar asociaciones al azar (117).

En general, es importante tener en cuenta que un hallazgo significativo en un estudio de asociación genética se puede interpretar como: i) una asociación directa, en el que el SNP identificado es la variante funcional que incluye el alelo de riesgo para la enfermedad; ii) una asociación indirecta, en la que el SNP identificado está en LD con la variante funcional y corresponde a un tagSNP; o iii) un falso positivo por efecto del azar, de

factores confusores como la subestructura poblacional y la mezcla genética o por un insuficiente control de calidad en los datos (142).

Conceptos relacionados con el control de calidad por SNP y por individuo en los estudios genéticos de asociación:

Según el tamaño de la muestra y el número de SNPs genotipados de manera simultánea con la plataforma seleccionada, es decir, si es a mediana o gran escala, se pueden modificar los umbrales de los controles de calidad de menos a más astringentes, respectivamente (117). Teniendo en cuenta lo anterior, a continuación se describen los conceptos relacionados con el control de calidad por SNP y por individuo.

- **Control de calidad por individuo.** Idealmente, este proceso debe incluir los siguientes cuatro pasos:
    a) *Identificación y eliminación de individuos con diferencias entre el sexo registrado y el sexo genético* (117). Este paso se basa en que los hombres tienen una sola copia del cromosoma X, heredada de su madre, mientras que las mujeres tienen dos copias heredadas de cada uno de sus progenitores. Debido a lo anterior, se espera que los hombres no sean **heterocigotos** para ningún marcador en este cromosoma, cuya característica se conoce como hemicigoto; mientras que las mujeres pueden ser **homocigotas** o heterocigotas para los marcadores en el cromosoma X. Entonces, desde el punto de vista genético, se considera que una muestra pertenece al sexo masculino cuando la tasa de homocigocidad en el cromosoma X es de 1.0, es decir para el 100% de los marcadores; por otro lado, se considera femenino una muestra cuya tasa de homocigocidad es de ≤ 0.2, es decir, hasta el 20% de los marcadores en el cromosoma X son homocigotos y el restante son heterocigotos. Con este paso es posible detectar cuales muestras no coinciden en el sexo registrado según la encuesta o la historia clínica con el sexo genético; lo cual puede indicar que hubo un error o cambio en las muestras genotipadas en el momento del montaje o contaminación o que se registró mal el sexo del individuo. Según el caso, es preciso determinar si se debe eliminar la muestra o si se puede rescatar.

b) *Identificación y eliminación de individuos con falla en el genotipado y con valores no esperados de tasa de heterocigocidad* (117). La baja calidad del DNA en una muestra, ya sea por degradación o por contaminación con proteínas u otras sustancias, puede favorecer a que ocurran fallas o interferencias en el genotipado; por lo anterior, se recomienda descartar las muestras en las cuales hay falla en el genotipado entre el 3% al 7% de los marcadores. Igualmente importante es descartar las muestras en las cuales exista una desviación considerable de la media de la tasa de heterocigocidad; por ejemplo, una tasa de heterocigocidad mucho menor a la media puede indicar endogamia, mientras que una tasa de heterocigocidad mayor puede indicar que hubo contaminación durante el procesamiento o montaje de las muestras.

c) *Identificación y eliminación de una muestra de cada par que se cataloguen como duplicadas o emparentadas* (117). Este paso es importante ya que en los estudios de asociación poblacionales se asume que cada individuo/muestra incluida, sea caso o control, representa una familia independiente. De esta manera, se consigue que los resultados del estudio sean un reflejo de la población general y no de familias específicas. Para identificar muestras duplicadas o emparentadas se debe calcular la proporción de alelos compartidos entre cada par de muestras, a nivel de autosomas y usando solo SNPs independientes (es decir que no estén en LD); esto es lo que se conoce como **identidad en estado, IBS**, en donde un IBS = 1 indica que el par de muestras comparten el 100% de los alelos y por tanto se consideran duplicadas. Con la información IBS es posible calcular si los alelos compartidos provienen de un mismo ancestro, lo que se conoce como **identidad por descendencia, IBD**; un IBD = 1 indica que el par de muestras son duplicados o que corresponden a gemelos monocigotos, un IBD = 0.5 indica que el 50% de los alelos entre el par de muestras provienen del mismo ancestro lo que indica que son parientes en primer grado, un IBD = 0.25 indica que son muestras emparentadas en segundo grado y un IBD = 0.125 indica que son muestras emparentadas en tercer grado. En estudios de asociación a gran escala, se recomienda eliminar la muestra de menor calidad entre cada par con un IBD = 0.1875, umbral que está entre los valores 0.25 - 0.125 mencionados.

d) *Identificación de estratificación poblacional* (117). Un posible confusor en los estudios genéticos de asociación poblacionales, es el sesgo en la recolección de casos y controles, ya que en este escenario las diferencias en frecuencias alélicas encontradas serán producto de diferencias en el origen de las muestras mas no relacionadas con el estatus de enfermedad, que es lo que se quiere estudiar. Mediante métodos como el escalamiento multidimensional es posible graficar las similitudes o distancias entre las muestras a analizar para concluir si los casos y controles son comparables o si provienen de diferentes fuentes.

- **Control de calidad por SNP.** Este proceso debe incluir los siguientes cuatro pasos:
  a) *Identificación y eliminación de SNPs con altas tasas de fallas en el genotipado* (117). Se recomienda eliminar los SNPs que hayan fallado en el genotipado en más del 5% de las muestras; esto con el fin de evitar asociaciones falsas en los análisis posteriores.
  b) *Identificación y eliminación de SNPs en desviación del* ***Equilibrio de Hardy-Weinberg, HWE*** (117). El HWE es un modelo que describe la relación entre las frecuencias alélicas y genotípicas en una población diploide sexualmente reproductiva con apareamiento al azar (69). Se considera que un marcador está en desequilibrio de HW cuando las frecuencias genotípicas observadas difieren significativamente de las esperadas, asumiendo una población en equilibrio, en la cual no ha ocurrido selección, ni mutaciones, ni migraciones y es lo suficientemente grande como para descartar efectos por **deriva génica** (69). Los SNPs en desequilibrio de HW se deben remover pues puede ser indicativo de errores en el genotipado; con el fin de evitar eliminar SNPs que puedan estar asociados con la enfermedad a estudio, este criterio solo debe evaluarse en los *"controles",* sin incluir los casos (117).
  c) *Identificación y eliminación de SNPs con diferencias significativas en las tasas de fallas en el genotipado entre casos y controles* (117). Este paso es importante para eliminar SNPs pobremente genotipados y evitar falsos-positivos en los análisis de asociación posteriores.
  d) *Eliminación de SNPs con MAF muy bajos* (117). Debido a la baja representación de alelos muy raros en la población general, se necesitarían tamaños de muestra demasiado grandes para encontrar asociaciones con la

enfermedad a estudio, por esta razón, la eliminación de estos SNPs no genera ningún impacto desfavorable en los estudios de asociación poblacionales.

Avances en la identificación de variantes genéticas comunes en el riesgo de CCR:

La teoría de *"enfermedad común, variantes comunes"* está respaldada por la evidencia actual sobre el papel de múltiples variantes genéticas comunes de susceptibilidad con efecto moderado o bajo en el desarrollo tumoral en colon y recto (15). Más aún, a partir de la creación del proyecto Internacional HapMap (140), se han implementado diversos estudios tipo GWAS en CCR los cuales han permitido la identificación de 42 **loci** con aproximadamente 63 variantes comunes de riesgo con una $P < 5 \times 10^{-8}$ (34-56); algunas de las cuales se han asociado también a adenomas colorrectales (ver **Tabla 16**) (44, 50, 147, 148).

Asumiendo que las variantes genéticas comunes fueran las únicas responsables del riesgo genético en CCR, *Tenesa A, et al.,* construyeron un modelo para estimar el número de variantes necesarias para alcanzar a explicar el total de la heredabilidad atribuida al CCR (hasta el ~35% de los casos) y concluyeron que alrededor de 170 variantes genéticas independientes, con MAF > 0.05, son necesarias para explicar el riesgo genético de CCR en población escocesa (15). Lo anterior sugiere que, a pesar de los esfuerzos, aún faltan por descubrir factores de susceptibilidad genética en CCR que contribuyan a su heredabilidad.

En la **Tabla 16**, se resumen los resultados de los SNPs que se han encontrado asociados al riesgo de CCR y PA. Como se mencionó antes, la mayoría de estos estudios a gran escala han sido conducidos en poblaciones europeas, asiáticas y blancos de Canadá y USA; sin embargo, también se han realizado esfuerzos en replicar los hallazgos en población negra de USA. En todos los casos para CCR la asociación obtenida en los estudios de replicación tuvo la misma dirección que en el descubrimiento, menos en el caso de la variante 10p14:rs10795668 (A;G) cuyo alelo de riesgo A se encontró como protector en población de raza blanca (52, 149), en Asiáticos (55, 150) y en judíos Azhkenazis (46), pero como factor de riesgo en población de raza negra de USA (151).

Recientemente se publicó un estudio tipo GWAS en población Hispana de USA, México, América Central y del Sur e Islas del Caribe (143), que reportó 17 variantes nuevas en 4 loci como *"sugestivos"* de riesgo para CCR en Hispanos; sin embargo, ninguno alcanzó significancia estadística después de la corrección con Bonferroni por pruebas múltiples (143). Este mismo estudio encontró que 20 de 58 variantes reportadas como asociadas al riesgo de CCR hasta el 2015, replicaron en población Hispana y la dirección de la asociación fue la misma que en las poblaciones donde se descubrieron (143); 16 de éstas variantes se muestran en la **Tabla 16**.

En concordancia con la vía de progresión adenoma-adenocarcinoma (13, 28, 29), *Carvajal-Carmona LG, et al.,* encontraron que las variantes 3q26.2:rs10936599 (C;T), 8q24.21:rs6983267 (G;T), 10p14:rs10795668 (A;G), 11q23:rs3802842 (A;C), 14q22.2:rs4444235 (C;T), 14q22.2:rs1957636 (A;G), 18q21.1:rs4939827 (C;T) y 20q12.3:rs961253 (A;C) se asociaron con el riesgo de adenomas en la misma dirección que con CCR (ver **Tabla 16**); sin embargo, otras variantes previamente asociadas con el riesgo de CCR, no fueron encontradas como asociadas al riesgo de adenomas (152); este hallazgo no es de extrañar, teniendo en cuenta que solo el 10% de los adenomas progresan a CCR (153), luego es de esperarse que existan variantes comunes y variantes únicas para cada caso.

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P* ALÉLICO REPORTADO	N		DESCUBRIMIENTO	REPLICACIÓN
						CASO	CONTROL		
8q24	-	CCR	Canadá, US, Europa	1.17 [1.12-1.23]	3.16×10^{-11}	7480	7779	Zanke BW et al., Nat Genet, 2007	
8q24.21 *(MYC)* / rs6983267 (G;T)	Alelo G	CCR	Reino Unido	1.27 [1.16-1.39] het 1.47 [1.34-1.62] hom 1.21 [1.15-1.27] alélico	1.27×10^{-14}	7954	6206	Tomlinson IP et al., Nat Genet, 2007	*US (raza blanca):* Berndt SI et al., Hum Mol Genet, 2008; Cicek MS et al., CEBP, 2009; He J et al., CEBP, 2011. *Asiáticos:* Matsuo K et al., BMC Cancer, 2009; Xiong F et al., CEBP, 2010; Cui R et al., Gut, 2011; Zhang B et al., Nat Genet, 2014. *US (raza negra):* He J et al., CEBP, 2011; Wang H et al., Hum Mol Genet, 2013. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Haiman et al., Nat Genet, 2007; S von Holst et al., BJC, 2010; Schmit SL et al., Carcinogenesis, 2014.
		PA		1.22 [1.10-1.34]	6.89×10^{-5}	1477	2136		
8q24 / rs7014346 (A;G)	Alelo A	CCR	Reino Unido, Alemania, Israel, España, Canadá y Japón	1.19 [1.15-1.23]	8.60×10^{-26}	14500	13294	Tenesa A et al., Nat Genet, 2008	*US (raza negra):* Kupfer SS et al. Gastroenterology, 2010. *Asiáticos:* Zhang B et al. Nat Genet, 2014. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al. Carcinogenesis, 2014.
8q24 / rs7837328 (A;G)	Alelo A	CCR	Japón	1.17 [1.10-1.24]	7.44×10^{-8}	6163	4494	Cui R et al., Gut, 2011	
18q21.1 *(SMAD7)* / rs4939827 (C;T)	Alelo C	CCR	Reino Unido	0.86 [0.79-0.92] het 0.73 [0.66-0.80] hom 0.85 [0.81-0.89] alélico	1.00×10^{-12}	7473	5984	Broderick P et al., Nat Genet, 2007	*Raza blanca:* Tenesa A et al., Nat Genet, 2008; Peters U et al., Hum Genet, 2012. *Asiáticos:* Xiong F et al., CEBP, 2010; Song Q et al., PLOS one, 2012; Zhang B et al., Nat Genet, 2014; Zhang B et al., Int J Cancer, 2014; Tenesa A et al., Nat Genet, 2008. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014.
18q21.1 *(SMAD7)* / rs7229639 (G;A)	Alelo A	CCR	China, Japon, Korea	1.22 [1.15-1.29] alélico	2.93×10^{-11}	4840	5925	Zhang B et al., Int J Cancer, 2014	*Asiáticos:* Zhang B et al., Nat Genet, 2014. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
15q13.3 - q14 *(GREM1)* / rs4779584 (C;T)	Alelo T	CCR	Reino Unido	1.23 [1.13-1.33] het 1.70 [1.41-2.04] hom 1.26 [1.19-1.34] alélico	4.44×10^{-14}	7961	6803	Jaeger E et al., Nat Genet, 2008	*Raza blanca:* He J et al., CEBP, 2011; Peters U et al., Hum Genet, 2012; Tomlinson et al., PLOS genetics, 2011. *Asiáticos:* Zhang B et al., Nat Genet, 2014.
15q13.3 *(GREM1)* / rs11632715 (A;G)	Alelo A	CCR	Reino Unido, US + Canadá (raza blanca), Australia, Finlandia, España	1.12 [1.08-1.16]	2.30×10^{-10}	13090	13884	Tomlinson IP et al., PLOS genetics, 2011	
15q13.3 *(GREM1)* / rs16969681 (C;T)	Alelo T	CCR	Reino Unido, US + Canadá (raza blanca), Australia, Finlandia, España	1.18 [1.11-1.25]	5.33×10^{-8}	12459	14061	Tomlinson IP et al., PLOS genetics, 2011	*US (raza negra):* Wang H et al., Hum Mol Genet, 2013. *Asiáticos:* Zhang B et al., Nat Genet, 2014.

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis (continuación…)

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P* ALÉLICO REPORTADO	N CASO	N CONTROL	DESCUBRIMIENTO	REPLICACIÓN
11q23 *(POU2AF1)* / rs3802842 (A;C)	Alelo C	CCR	Reino Unido, Alemania, Israel, España, Canadá y Japón	1.11 [1.08-1.15]	5.82×10^{-10}	14500	13294	Tenesa A et al., Nat Genet, 2008	*Raza blanca:* He J et al., CEBP, 2011; Peters U et al., Hum Genet, 2012. *Asiáticos:* Zhang B et al., Nat Genet, 2014. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014.
14q22.2 *(BMP4)* / rs1957636 (A;G)	Alelo A	CCR	Reino Unido, US + Canadá (raza blanca), Australia, Finlandia, España	1.08 [1.06-1.11]	1.36×10^{-9}	24487	23722	Tomlinson IP et al., PLOS genetics, 2011	***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
14q22.2 *(BMP4)* / rs4444235 (C;T)	Alelo C	CCR	Reino Unido, Canadá, Alemania, Finlandia	1.11 [1.08-1.15]	8.1×10^{-10}	20186	20855	COGENT Study et al., Nat Genet, 2008	*Raza blanca:* Tomlinson IP et al., PLOS genetics, 2011 *Asiáticos:* Zhang B et al., Nat Genet, 2014
16q22.1 *(CDH1)* / rs9929218 (A;G)	Alelo A	CCR	Reino Unido, Canadá, Alemania, Finlandia	0.91 [0.89-0.94]	1.2×10^{-8}	20186	20855	COGENT Study et al., Nat Genet, 2008	*Asiáticos:* Zhang B et al., Nat Genet, 2014
19q13.1 *(RHPN2)* / rs10411210 (C;T)	Alelo T	CCR	Reino Unido, Canadá, Alemania, Finlandia	0.87 [0.83-0.91]	4.6×10^{-9}	20186	20855	COGENT Study et al., Nat Genet, 2008	*Asiáticos:* Zhang B et al., Nat Genet, 2014 ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
20q12.3 *(BMP2)* / rs961253 (A;C)	Alelo A	CCR	Reino Unido, Canadá, Alemania, Finlandia	1.12 [1.08-1.16]	2.0×10^{-10}	20186	20855	COGENT Study et al., Nat Genet, 2008	*US (raza negra):* He Jet al., CEBP, 2011; Wang H et al., Hum Mol Genet, 2013. *Raza blanca:* Tomlinson IP et al., PLOS genetics, 2011 ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014
20q12.3 *(BMP2)* / rs4813802 (G;T)	Alelo G	CCR	Reino Unido, US + Canadá (raza blanca), Australia, Finlandia, España	1.09 [1.06-1.12]	7.52×10^{-11}	24357	23496	Tomlinson IP et al., PLOS genetics, 2011	***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014
8q23.3 *(EIF3H)* / rs16892766 (A;C)	Alelo C	CCR	Reino Unido, Alemania, Finlandia, Holanda, Australia	1.27 [1.20-1.34] het 1.43 [1.13-1.82] hom 1.25 [1.19-1.32] alélico	3.3×10^{-18}	10731	10961	Tomlinson IP et al., Nat Genet, 2008	*Raza blanca:* S von Holst et al., BJC, 2010; Peters U et al., Hum Genet, 2012. *US (raza negra):* Wang H et al., Hum Mol Genet, 2013. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014; Liao M et al., Scientific Reports, 2015.
10p14 / rs10795668 (A;G)	Alelo A	CCR	Reino Unido, Alemania, Finlandia, Holanda, Australia	0.87 [0.83-0.91] het 0.80 [0.74-0.86] hom 0.89 [0.86-0.91] alélico	2.5×10^{-13}	10731	10961	Tomlinson IP et al., Nat Genet, 2008	*Raza blanca:* S von Holst et al., BJC, 2010. *US (raza negra):* Kupfer SS et al., Gastroenterology, 2010*. *Asiáticos:* Qin Q et al., PLOS one, 2013; Zhang B et al., Nat Genet, 2014. ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.** *Otros:* Schmit SL et al., Carcinogenesis, 2014.

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis (continuación…)

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P* ALÉLICO REPORTADO	N		DESCUBRIMIENTO	REPLICACIÓN
						CASO	CONTROL		
1q41 *(DUSP10)* / rs6687758 (A;G)	Alelo G	CCR	Reino Unido	1.09 [1.06-1.12]	2.27×10^{-9}	18095	20197	Houlston RS et al., Nat Genet, 2010	*Asiáticos:* Zhang B et al., Nat Genet, 2014.
1q41 *(DUSP10)* / rs6691170 (G;T)	Alelo T	CCR	Reino Unido	1.06 [1.03-1.09]	9.55×10^{-10}	18095	20197	Houlston RS et al., Nat Genet, 2010	
12q13.13 *(DIP2B)* / rs11169552 (C;T)	Alelo T	CCR	Reino Unido	0.92 [0.90-0.95]	1.89×10^{-10}	18095	20197	Houlston RS et al., Nat Genet, 2010	*Asiáticos:* Zhang B et al., Nat Genet, 2014. *Otros:* Schmit SL et al., Carcinogenesis, 2014.
12q13.13 *(DIP2B)* / rs7136702 (C;T)	Alelo T	CCR	Reino Unido	1.06 [1.04-1.08]	4.02×10^{-8}	18095	20197	Houlston RS et al., Nat Genet, 2010	
20q13.33 *(LAMA)* / rs4925386 (C;T)	Alelo T	CCR	Reino Unido	0.93 [0.91-0.95]	1.89×10^{-10}	18095	20197	Houlston RS et al., Nat Genet, 2010	*Otros:* Schmit SL et al., Carcinogenesis, 2014.
3q26.2 *(TERC)* / rs10936599 (C;T)	Alelo T	CCR	Reino Unido	0.93 [0.91-0.96]	3.39×10^{-8}	18095	20197	Houlston RS et al., Nat Genet, 2010	*Asiáticos:* Zhang B et al., Nat Genet, 2014.
6q26-q27 *(SLC22A3)* / rs7758229 (G;T)	Alelo T	CCR	Asia (Japón-Korea)	1.28 [1.18-1.39]	7.92×10^{-9}	2137	5506	Cui R et al., Gut, 2011	*US (raza negra):* Wang H et al., Hum Mol Genet, 2013
5p15.33 *(TERT-CLPTM1L)* / rs2736100 (G;T)	Alelo T	CCR	Reino Unido	1.07 [1.04-1.11]	2.49×10^{-5} (*P* Bonferroni 1.82×10^{-3})	16039	16430	Kinnersley B et al., BJC, 2012	
5q31.1 *(PITX1)* / rs647161 (A;C)	Alelo A	CCR	Asia (China, Korea, Japón)	1.17 [1.11-1.22)	3.77×10^{-10}	7315	11564	Jia WH et al., Nat Genet, 2013	*Raza blanca:* Jia WH et al., Nat Genet, 2013 *US (raza negra):* Wang H et al., Hum Mol Genet, 2013 *Asiáticos:* Zhang B et al., Nat Genet, 2014
20p12.3 *(HAO1)* / rs2423279 (C;T)	Alelo C	CCR	Asia (China, Korea, Japón)	1.14 [1.08-1.19]	2.29×10^{-7}	7325	11560	Jia WH et al., Nat Genet, 2013	*Raza blanca:* Jia WH et al., Nat Genet, 2013 *Asiáticos:* Zhang B et al., Nat Genet, 2014 ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
10q26.12 *(HSPA12A)* / rs1665650 (T;C)	Alelo T	CCR	Asia (China, Korea, Japón)	1.13 [1.08-1.19]	$8.58 \times 10e^{-7}$	7290	11557	Jia WH et al., Nat Genet, 2013	
12p13.32 *(CCND2)* / rs10774214 (T;C)	Alelo T	CCR	Asia (China, Korea, Japón)	1.17 [1.11-1.23]	5.48×10^{-10}	7295	11546	Jia WH et al., Nat Genet, 2013	*Raza blanca:* Jia WH et al., Nat Genet, 2013 *Asiáticos:* Zhang B et al., Nat Genet, 2014
12p13.32 *(CCND2)* / rs3217810 (C;T)	Alelo T	CCR + PA	Ancestría Europea y Asiática	1.20 [1.12-1.28]	5.86×10^{-8}	13654	16022	Peters U et al., Gastroenterology, 2013	*Raza blanca:* Whiffin N et al., Hum Mol Genet, 2014 ***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
12p13.32 *(CCND2)* / rs3217901 (A;G)	Alelo G	CCR + PA	Ancestria Europea y Asiática	1.10 [1.06-1.14]	3.31×10^{-7}	15752	21771	Peters U et al., Gastroenterology, 2013	

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis (continuación…)

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P*ALÉLICO REPORTADO	N		DESCUBRIMIENTO	REPLICACIÓN
						CASO	CONTROL		
2q32.3 *(NABP1)* / rs11903757 (C;T)	Alelo C	CCR + PA	Ancestria Europea y Asiática	1.16 [1.10-1.22]	3.71×10^{-8}	15752	21771	Peters U et al., Gastroenterology, 2013	*Asiáticos:* Zhang B et al., Nat Genet, 2014. *Otros:* Schmit SL et al., Carcinogenesis, 2014.
12q24.21 *(TBX3)* / rs59336 (A;T)	Alelo T	CCR + PA	Ancestria Europea y Asiática	1.09 [1.06-1.13]	3.76×10^{-7}	15752	21771	Peters U et al., Gastroenterology, 2013	
1q25.3 *(LAMC1)* / rs10911251 (A;C)	Alelo A	CCR + PA	Ancestria Europea y Asiática	1.09 [1.06-1.13]	9.45×10^{-8}	15752	21771	Peters U et al., Gastroenterology, 2013	*Raza blanca:* Whiffin N et al., Hum Mol Genet, 2014
1p33 *(SLC5A9)* / rs12080929 (C;T)	Alelo C	CCR	Sur de España	0.86 [0.78-0.96]	5.52×10^{-3} (metaanálisis)	2362	2517	Fernandez-Rozadilla et al. BMC Genomics, 2013	
8p12 *(DUSP4)* / rs11987193 (C;T)	Alelo C	CCR	Sur de España	0.79 [0.71-0.88]	6.98×10^{-5} (metaanálisis)	2362	2517	Fernandez-Rozadilla et al. BMC Genomics, 2013	
21q22.2 *(IGSF5)* / rs2837156 (G;A)	Alelo G	PA	Ancestria Europea	2.22 [1.62-3.03]	3.2×10^{-7}	139	1267	Wang J et al., Clin Cancer Res, 2013	
21q22.2 *(IGSF5)* / rs7278863 (A;G)	Alelo A	PA	Ancestria Europea	2.48 [1.80-3.42]	1.4×10^{-8}	139	1267	Wang J et al., Clin Cancer Res, 2013	
21q22.2 / rs2837237 (G;A)	Alelo G	PA	Ancestria Europea	2.48 [1.82-3.38]	3.6×10^{-9}	139	1267	Wang J et al., Clin Cancer Res, 2013	
21q22.2 / rs2837241 (A;C)	Alelo A	PA	Ancestria Europea	2.48 [1.82-3.38]	3.6×10^{-9}	139	1267	Wang J et al., Clin Cancer Res, 2013	
21q22.2 / rs2837254 (A;G)	Alelo A	PA	Ancestria Europea	2.55 [1.86-3.51]	2.9×10^{-9}	139	1267	Wang J et al., Clin Cancer Res, 2013	
21q22.2 / rs741864 (A;G)	Alelo A	PA	Ancestria Europea	2.48 [1.80-3.41]	1.1×10^{-8}	139	1267	Wang J et al., Clin Cancer Res, 2013	
3p24.1 / rs1381392 (A;G)	Alelo A	PA	Ancestria Europea	2.01 [1.52-2.65]	7.4×10^{-7}	139	1267	Wang J et al., Clin Cancer Res, 2013	
3p24.1 / rs17651822 (A;G)	Alelo A	PA	Ancestria Europea	2.16 [1.61-2.91]	2.1×10^{-7}	139	1267	Wang J et al., Clin Cancer Res, 2013	
13q33.2 / rs1535989 (G;A)	Alelo G	PA	Ancestria Europea	2.09 [1.50-2.91]	8.9×10^{-6}	139	1267	Wang J et al., Clin Cancer Res, 2013	
10q24.2 / rs1035209 (C;T)	Alelo T	CCR	Reino Unido, US + Canadá (raza blanca), Australia, Suecia	1.12 [1.08-1.16]	4.54×10^{-11}	14037	15937	Whiffin N et al., Hum Mol Genet, 2014	

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis (continuación…)

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P* ALÉLICO REPORTADO	N		DESCUBRIMIENTO	REPLICACIÓN
						CASO	CONTROL		
10q22.3 *(ZMIZ1-AS1)* / rs704017 (G;A)	Alelo G	CCR	China, Korea del Sur y Japón	1.10 [1.06-1.13]	2.07×10^{-8}	14963	31945	Zhang B et al., Nat Genet, 2014	*Raza blanca:* Zhang B et al., Nat Genet, 2014
10q25.2 *(TCF7L2)* / rs11196172 (A;G)	Alelo A	CCR	China, Korea del Sur y Japón	1.14 [1.10-1.18]	1.04×10^{-12}	14963	31945	Zhang B et al., Nat Genet, 2014	
11q12.2 *(MYRF)* / rs174537(G;T)	Alelo G	CCR	China, Korea del Sur y Japón	1.16 [1.12-1.19]	9.22×10^{-21}	14963	31945	Zhang B et al., Nat Genet, 2014	*Raza blanca:* Zhang B et al., Nat Genet, 2014
11q12.2 *(FEN1)* / rs4246215 (G;T)	Alelo G	CCR	China, Korea del Sur y Japón	1.15 [1.12-1.19]	7.65×10^{-20}	14963	31945	Zhang B et al., Nat Genet, 2014	*Raza blanca:* Zhang B et al., Nat Genet, 2014
11q12.2 *(FADS1)* / rs174550 (T;C)	Alelo T	CCR	China, Korea del Sur y Japón	1.15 [1.12-1.19]	1.58×10^{-19}	14963	31945	Zhang B et al., Nat Genet, 2014	*Raza blanca:* Zhang B et al., Nat Genet, 2014
11q12.2 *(FADS2)* / rs1535 (A;G)	Alelo A	CCR	China, Korea del Sur y Japón	1.15 [1.12-1.19]	8.21×10^{-20}	14963	31945	Zhang B et al., Nat Genet, 2014	*Raza blanca:* Zhang B et al., Nat Genet, 2014
12p13.31 *(CD9)* / rs10849432 (T;C)	Alelo T	CCR	China, Korea del Sur y Japón	1.14 [1.09-1.18]	5.81×10^{-10}	14963	31945	Zhang B et al., Nat Genet, 2014	
17p13.3 *(NXN)* / rs12603526 (C;T)	Alelo C	CCR	China, Korea del Sur y Japón	1.10 [1.06-1.14]	3.42×10^{-8}	14963	31945	Zhang B et al., Nat Genet, 2014	
19q13.2 *(TGFB1)* / rs1800469 (G;A)	Alelo G	CCR	China, Korea del Sur y Japón	1.09 [1.06-1.12]	1.17×10^{-8}	14963	31945	Zhang B et al., Nat Genet, 2014	
19q13.2 *(B9D2)* / rs2241714 (C;T)	Alelo C	CCR	China, Korea del Sur y Japón	1.09 [1.06-1.12]	1.36×10^{-8}			Zhang B et al., Nat Genet, 2014	
4q26 *(TRAM1L1-NDST3)* / rs3987 (A;G)	Alelo C	CCR	España	1.36 [-]	4.02×10^{-8}	1224	1882	Real LM et al., PLOS one, 2014	
4q32.2 *(FSTL5)* / rs35509282 (A;T)	Alelo A	CCR	Israel (Ashkenazi)	1.53 ± 0.07 SE	8.2×10^{-9}	3593	2328	Schmit SL et al., Carcinogenesis, 2014	
5q23.3 *(CDC42SE2-CHSY3)* / rs12522693 (G;A)	Alelo A	CCR	China	1.31 [1.19-1.45]	2.08×10^{-8}	3633	4678	Jiang K et al., Oncotarget, 2015	
5q23.3 *(CDC42SE2-CHSY3)* / rs10035791 (G;A)	Alelo A	CCR	China	1.25 [1.14-1.36]	8.20×10^{-7}	3627	4656	Jiang K et al., Oncotarget, 2015	
5q23.3 *(CDC42SE2-CHSY3)* / rs80007597 (G;C)	Alelo C	CCR	China	1.29 [1.17-1.42]	3.63×10^{-7}	3620	4655	Jiang K et al., Oncotarget, 2015	
17q12 *(ASIC2-CCL2)* / rs17836917 (G;A)	Alelo A	CCR	China	0.75 [0.68-0.83]	4.55×10^{-8}	3634	4678	Jiang K et al., Oncotarget, 2015	

Tabla 16 Locus / SNPs reportados como asociados a CCR o PA en estudios tipo GWAS o meta análisis (continuación…)

LOCUS / SNP	ALELO DE RIESGO	FENOTIPO	POBLACIÓN	OR [95%CI] REPORTADO	*P* ALÉLICO REPORTADO	N		DESCUBRIMIENTO	REPLICACIÓN
						CASO	CONTROL		
10q25 *(VTI1A)* / rs12241008 (C;T)	Alelo C	CCR	US (raza negra), Japón, Europa	1.13 [1.09-1.18]	1.4×10^{-9} (metaanálisis)	26710	21343	Wang H et al., Nat Commun, 2015	***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
1p36.12 *(WNT4-CDC42)* / rs72647484 (T;C)	Alelo T	CCR	Ancestria Europea	1.24 [1.15-1.33]	1.21×10^{-8}	7577	9979	Al-Tassan NA et al., Scientific Reports, 2015	
16q24.1 *(FOXL1)* / rs16941835 (G;C)	Alelo C	CCR	Ancestria Europea	1.16 [1.09-1.22]	5.06×10^{-8}	7577	9979	Al-Tassan NA et al., Scientific Reports, 2015	
3p22.1 / rs35360328 (A;T)	Alelo A	CCR	Ancestria Europea y Asiática	1.14 [1.09–1.19]	3.1×10^{-9}	23024	29625	Schumacher FR et al., Nat Commun,2015	
3p14.1 / rs812481 (G;C)	Alelo G	CCR	Ancestria Europea y Asiática	1.09 [1.05–1.11]	2.0×10^{-8}	23024	29625	Schumacher FR et al., Nat Commun,2015	
10q24.2 / rs11190164 (G;A)	Alelo G	CCR	Ancestria Europea y Asiática	1.09 [1.06–1.12]	4.0×10^{-8}	23024	29625	Schumacher FR et al., Nat Commun,2015	
12q24.12 / rs3184504 (C;T)	Alelo C	CCR	Ancestria Europea y Asiática	1.09 [1.06–1.12]	1.7×10^{-8}	23024	29625	Schumacher FR et al., Nat Commun,2015	
12q24.22 / rs73208120 (G;T)	Alelo G	CCR	Ancestria Europea y Asiática	1.16 [1.11–1.23]	2.8×10^{-8}	23024	29625	Schumacher FR et al., Nat Commun,2015	
20q13.13 / rs6066825 (A;G)	Alelo A	CCR	Ancestria Europea y Asiática	1.09 [1.06–1.12]	4.4×10^{-9}	23024	29625	Schumacher FR et al., Nat Commun,2015	
14q23.1 *(RTN1)* / rs17094983 (A;G)	Alelo A	CCR	Ancestria Europea	0.87 [0.83-0.91]	2.5×10^{-10}	16517	14487	Lemire M et al., Hum Genet, 2015	*Raza negra:* Lemire M et al., Hum Genet, 2015
11q13.4 *(POLD3)* / rs3824999 (G;T)	Alelo G	CCR	Reino Unido, Finlandia, Suecia, Croacia, Japón	1.08 [1.05-1.10]	3.65x10e-10	21096	19555	Dunlop MG et al., Nat Genet, 2012	*Otros:* Schmit SL et al., Carcinogenesis, 2014.
6q21 *(CDKN1A)* / rs1321311 (A;C)	Alelo A	CCR	Reino Unido, Finlandia, Suecia, Croacia, Japón	1.10 [1.07-1.13]	1.14x10e-10	21096	19555	Dunlop MG et al., Nat Genet, 2012	***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**
Xp22.2 *(SHROOM2)* / rs5934683(T;C)	Alelo T	CCR	Reino Unido, Finlandia, Suecia, Croacia, Japón	1.07 [1.04-1.10]	7.30x10e-10	21096	19555	Dunlop MG et al., Nat Genet, 2012	***Hispanos:*** **Schmit SL et al., Carcinogenesis, 2016.**

Se registra el OR correspondiente al modelo alélico, usando el alelo de riesgo; lo anterior, salvo se especifique otro modelo, por ejemplo: Het, heterocigoto; Hom, homocigoto.

* Asociación inversa a la encontrada inicialmente

SE, error estándar; US, Estados Unidos

Vías de señalización implicadas en el riesgo de CCR, sugeridas a partir de estudios de asociación:

El estudio de marcadores genéticos asociados al riesgo de desarrollar enfermedades complejas como el cáncer, ha sido de gran interés para la comunidad científica, no solo con el fin de aportar al entendimiento de las bases biológicas determinantes de la variación en los fenotipos, ejemplo: lesiones preneoplásicas versus neoplasias colorrectales (152), sino también para encontrar variantes genéticas involucradas en el desarrollo de una enfermedad que sirvan: i) como marcadores para la estratificación de individuos con alto riesgo de desarrollar la enfermedad y ii) como guías en la identificación de vías de señalización que puedan ser **blancos terapéuticos** (140, 141).

Como se puede observar en la **Tabla 16**, muchas de las variantes encontradas como asociadas a tumores colorrectales en estudios tipo GWAS se ubican cerca de genes que codifican proteínas que participan en vías de señalización posiblemente implicadas en el desarrollo o progresión de neoplasias colorrectales. Por ejemplo, se ha evidenciado un aumento en la unión del alelo de riesgo 8q24.21:rs6983267 (G) con el factor de transcripción TCF7L2, el principal efector transcripcional de la vía Wnt, que es activado por B-catenina una vez que éste se transloca al núcleo; vale la pena resaltar que el protooncogén *MYC* es uno de los blancos de ésta vía de señalización, la cual ya se ha encontrado relacionada con el desarrollo de cánceres gastrointestinales (154-156). Si bien se ha observado que este locus interactúa con el promotor de *MYC*, no se ha demostrado un aumento en la expresión de *MYC* ligado al alelo 8q24.21:rs6983267 (G), por lo cual se cree que el efecto de este imbalance es entonces a largo plazo (156). Otro grupo encontró que el mecanismo mediante el cual existe una sobre representación del alelo de riesgo 8q24.21:rs6983267 (G) en los CCR se debe a la amplificación del haplotipo que contiene este alelo (157, 158).

El SNP 15q13.3:rs16969681 (C;T) está ubicado en una región con modificación activa de cromatina que actúa como **enhancer** específico de alelo y tejido (159). El factor de transcripción específico del intestino, CDX2, y TCF7L2 se unen cerca a este SNP especialmente en presencia del alelo de riesgo T, lo que a su vez se asoció a una mayor expresión del gen *GREM1*, el cual se encontró que favorece la incidencia de tumores

intestinales en ratones *Apc^{min}* (159). *GREM1* codifica la proteína gremlin 1 que se expresa principalmente en las **células mesenquimales** de la base de las **criptas intestinales** y actúa como antagonista de las proteínas BMP, supresores de tumor que hacen parte de la vía TGF beta y que además actúan como inhibidores de la vía de señalización Wnt; de esta manera, la expresión de gremlin 1 favorece un fenotipo de **transición epitelial-mesenquimal** al permitir la activación de la vía Wnt (160).

La longitud de los telómeros se mantiene gracias a la actividad de la enzima telomerasa en las células altamente proliferativas; esta enzima consta de una subunidad catalítica con actividad de transcriptasa reversa conocida como hTERT, la cual usa como molde una molécula de RNA llamada hTERC, para elongar los telómeros y favorecer de esta manera la proliferación celular (161). Las células cancerosas son altamente proliferativas y en parte se ha evidenciado que esta capacidad es mantenida por la sobreexpresión de la enzima telomerasa (161). *Jones AM, et al.,* encontraron que el alelo mayor del SNP 3q26.2:rs10936599 (C) en el gen *hTERC*, se asoció no solo a un incremento en el riesgo de CCR (38, 55) y PA (152), sino que también se asoció a un aumento en la longitud de los telómeros en linfocitos periféricos de pacientes con CCR, por lo cual este estudio concluye que los telómeros largos son un factor de riesgo para CCR al simular un fenotipo de células madre hiperproliferativas (161).

2.1 Replicación de variantes genéticas previamente asociadas con CCR, en colombianos incluidos en el estudio

Teniendo en cuenta que para cuando se terminó el periodo de recolección de muestras del estudio, en el año 2011, habían reportadas alrededor de 16 variantes genéticas asociadas al riesgo de CCR en poblaciones europeas, se realizaron ensayos de TaqMan en 959 muestras colombianas. Un total de 14 variantes fueron genotipadas con una tasa de 0.98.

El propósito de esta tercera parte del estudio, es replicar algunas variantes genéticas previamente asociadas con CCR, en la muestra de casos y controles colombianos del estudio.

2.1.1 Objetivos específicos

2.1.1.1 Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, de 14 SNPs previamente publicados, mediante análisis básicos de asociación por SNP *(X^2)* y regresiones logísticas ajustadas.

2.1.2 Métodos

Alrededor de ~ 900 muestras, entre casos y controles, fueron genotipadas por TaqMan para 14 variantes seleccionadas con una tasa de genotipado de 0.98. Se compararon las frecuencias alélicas entre los grupos del estudio por medio de análisis básicos por SNP *(X^2)* para evaluar diferencias significativas. Adicionalmente, se corrieron regresiones

logísticas ajustadas por edad, sexo y las ancestrías globales europea y africana. Nuevamente, la variable *"Array"* se incluyó en los modelos para corregir los análisis según la base de datos con la que se calcularon las ancestrías. *Ver el Anexo 1 - Materiales y métodos en análisis genéticos, para obtener información más detallada de los métodos.*

2.1.3 Resultados

Replicación de 14 SNPs publicados como asociados con CCR en poblaciones europeas, en la muestra de casos y controles colombianos del estudio

En la **Tabla 17** y **Tabla 18** se muestran los resultados obtenidos en los análisis de asociación de 14 SNPs publicados como de riesgo para CCR en poblaciones europeas hasta el año 2011. Todos los SNPs, salvo el 10p14:rs10795668 y el 14q22.2:rs4444235, se encontraron en HWE en controles ($P > 0.05$).

Tabla 17 Replicación de SNPs previamente reportados, en el riesgo de PA en colombianos

LOCUS / SNP	FRECUENCIA DEL ALELO MENOR POR FENOTIPO			ANÁLISIS BÁSICO DE ASOCIACIÓN POR SNP (X^2)			REGRESIÓN LOGÍSTICA AJUSTADA POR EDAD + SEXO + ANCESTRÍAS (EUROPEA Y AFRICANA)*		
	ALELO MENOR	PA	CONTROLES	OR	[95% CI]	*P*	OR	[95% CI]	*P*
8q24.21 / rs10505477 (G;A)	G	0.431	0.463	0.88	[0.67-1.15]	0.35	0.93	[0.70-1.24]	0.63
8q24.21 / rs6983267 (G;T)	T	0.399	0.429	0.89	[0.68-1.16]	0.39	0.96	[0.72-1.28]	0.76
8q24 / rs7014346 (A;G)	A	0.318	0.299	1.09	[0.82-1.46]	0.54	1.04	[0.77-1.42]	0.79
10p14 / rs10795668 (A;G)	A	0.271	0.297	0.88	[0.65-1.20]	0.42	0.82	[0.59-1.13]	0.22
11q23 / rs3802842 (A;C)	C	0.241	0.248	0.96	[0.70-1.31]	0.81	0.98	[0.71-1.36]	0.92
14q22.2 / rs1957636 (T;C)	C	0.452	0.454	0.99	[0.76-1.30]	0.96	0.94	[0.70-1.25]	0.66
14q22.2 / rs4444235 (C;T)	C	0.427	0.447	0.92	[0.70-1.21]	0.56	1.00	[0.73-1.37]	1.00
15q13.3 / rs11632715 (A;G)	G	0.442	0.444	0.99	[0.76-1.30]	0.96	1.06	[0.79-1.42]	0.69
15q13.3 - q14 / rs4779584 (C;T)	T	0.331	0.289	1.22	[0.91-1.64]	0.18	1.21	[0.87-1.68]	0.26
16q22.1 / rs9929218 (A;G)	A	0.203	0.213	0.94	[0.67-1.30]	0.69	1.00	[0.70-1.42]	0.98
18q21.1 / rs4939827 (C;T)	T	0.337	0.343	0.97	[0.73-1.29]	0.83	0.96	[0.71-1.30]	0.80
19q13.1 / rs10411210 (C;T)	T	0.186	0.186	1.00	[0.71-1.42]	1.00	1.04	[0.73-1.50]	0.82
20q12.3 / rs4813802 (G;T)	G	0.385	0.416	0.88	[0.66-1.16]	0.36	0.92	[0.69-1.24]	0.60
20q12.3 / rs961253 (A;C)	A	0.332	0.351	0.92	[0.69-1.23]	0.57	0.83	[0.60-1.15]	0.26

* Adicionalmente ajustado por *"Array"*

Ninguno de los SNPs se observó asociado al riesgo de PA en los análisis básicos por SNP *(X^2)* ni en las regresiones logísticas ajustadas (ver **Tabla 17**).

Con respecto al CCR, se observó un efecto protector del alelo menor del SNP 11q23:rs3802842 (C) al compararlo con los controles en los análisis básicos de asociación por SNP *(X^2)* (0.198 versus 0.248; OR 0.75; 95%CI 0.57-0.99; *P* = 0.04); sin embargo, en los análisis ajustados la significancia fue marginal (*P* = 0.08) (ver **Tabla 18**).

Tabla 18 Replicación de SNPs previamente reportados, en el riesgo de CCR en colombianos

LOCUS / SNP	FRECUENCIA DEL ALELO MENOR POR FENOTIPO			ANÁLISIS BÁSICO DE ASOCIACIÓN POR SNP *(X^2)*			REGRESIÓN LOGÍSTICA AJUSTADA POR EDAD + SEXO + ANCESTRÍAS (EUROPEA Y AFRICANA)*		
	ALELO MENOR	CCR	CONTROLES	OR	[95% CI]	*P*	OR	[95% CI]	*P*
8q24.21 / rs10505477 (G;A)	G	0.435	0.463	0.90	[0.71-1.13]	0.34	0.95	[0.75-1.20]	0.64
8q24.21 / rs6983267 (G;T)	T	0.424	0.429	0.98	[0.78-1.24]	0.88	1.04	[0.82-1.31]	0.75
8q24 / rs7014346 (A;G)	A	0.346	0.299	1.25	[0.98-1.59]	0.08	1.27	[0.98-1.64]	0.07
10p14 / rs10795668 (A;G)	A	0.282	0.297	0.93	[0.72-1.20]	0.58	0.91	[0.70-1.18]	0.46
11q23 / rs3802842 (A;C)	C	0.198	0.248	*0.75*	*[0.57-0.99]*	***0.04***	0.78	[0.59-1.03]	0.08
14q22.2 / rs1957636 (T;C)	C	0.423	0.454	0.88	[0.70-1.11]	0.29	0.90	[0.72-1.14]	0.40
14q22.2 / rs4444235 (C;T)	C	0.446	0.447	0.99	[0.79-1.25]	0.96	1.03	[0.81-1.32]	0.80
15q13.3 / rs11632715 (A;G)	G	0.418	0.444	0.90	[0.71-1.13]	0.36	0.90	[0.71-1.14]	0.38
15q13.3 - q14 / rs4779584 (C;T)	T	0.322	0.289	1.17	[0.91-1.51]	0.21	1.11	[0.85-1.43]	0.44
16q22.1 / rs9929218 (A;G)	A	0.196	0.213	0.90	[0.68-1.19]	0.46	0.87	[0.66-1.17]	0.36
18q21.1 / rs4939827 (C;T)	T	0.334	0.343	0.96	[0.76-1.22]	0.73	0.96	[0.75-1.23]	0.75
19q13.1 / rs10411210 (C;T)	T	0.161	0.186	0.84	[0.62-1.14]	0.26	0.83	[0.61-1.13]	0.24
20q12.3 / rs4813802 (G;T)	G	0.365	0.416	0.81	[0.64-1.02]	0.08	0.84	[0.66-1.07]	0.16
20q12.3 / rs961253 (A;C)	A	0.381	0.351	1.14	[0.90-1.44]	0.29	1.10	[0.86-1.42]	0.45

* Adicionalmente ajustado por *"Array"*
Valores en *itálica* y subrayados hacen referencia a que la asociación observada en colombianos es inversa a la reportada en europeos

Por otro lado, se observó una tendencia en la asociación del alelo menor del SNP 8q24:rs7014346 (A) como factor de riesgo en el desarrollo de CCR comparado con controles, tanto en los análisis básicos de asociación (0.346 versus 0.299; *P* = 0.08), como en los análisis ajustados (OR 1.27; 95%CI 0.98-1.64; *P* = 0.07) (ver **Tabla 18**). Finalmente, se observó una significancia marginal del efecto protector para CCR del alelo menor (G) del SNP 20q12.3:rs4813802 en los análisis básicos de asociación (*P* = 0.08), pero no en los análisis ajustados (*P* = 0.16).

2.2 Estudio de la asociación de variantes genéticas comunes, no antes reportadas, en el riesgo de tumores colorrectales en colombianos

Teniendo en cuenta la falta de conocimiento sobre la heredabilidad en CCR y que los estudios tipo GWAS realizados hasta el momento se han desarrollado principalmente en poblaciones de ascendencia europea y asiática, se hace importante la conducción de estudios en poblaciones Latinas, que por su estructura genética mixta tricontinental comprende una gran oportunidad para la identificación de nuevas variantes genéticas asociadas al riesgo de la enfermedad.

El propósito de esta cuarta parte del estudio, es descubrir nuevas variantes genéticas comunes, asociadas al riesgo de PA y CCR en colombianos, tomando ventaja del alto grado de mezcla de nuestra población.

2.2.1 Objetivos específicos

2.2.1.1 Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, usando una aproximación de genes candidatos (CG)

2.2.1.2 Evaluar diferencias significativas en las frecuencias alélicas entre PA o CCR comparado con controles colombianos, usando una aproximación de genoma completo (GWAS)

2.2.2 Métodos

Los análisis de asociación de genes candidatos se corrieron en las muestras genotipadas con la plataforma CG limpia, la cual incluye los genotipos de 1237 SNPs en 483 muestras colombianas (ver **Figura - Anexos B** y **Tabla - Anexos A**). Por otro lado, los análisis de asociación de genoma completo se realizaron en las muestras genotipadas con la plataforma GWAS limpia, incluyendo solo marcadores en autosomas y cromosoma X (719571 SNPs en 415 muestras colombianas) (ver **Figura - Anexos D** y **Tabla - Anexos A**).

Por separado con cada aproximación, CG y GWAS, se realizaron análisis de asociación básicos por SNP usando la prueba de χ^2 para evaluar diferencias en MAF entre casos y controles. *Ver el Anexo 1 - Materiales y métodos en análisis genéticos, para obtener información más detallada de los métodos.*

Con la base de datos CG se corrieron regresiones logísticas de todos SNPs incluidos, ajustando por edad, ancestría global europea, ancestría global africana y *"Array"*; este último porque para las 85 muestras sobrelapadas se usaron las ancestrías obtenidas con GWAS y para las demás se usaron las obtenidas con la plataforma CG. En este modelo no se corrigió por sexo, puesto que al hacerlo resultaban NAs. Se tomó como valor significativo una $P < 0.05$ corregida por Bonferroni (es decir, 0.05 / 1237 = P nominal < 4.04×10^{-5}). Adicionalmente, se realizaron análisis confirmatorios por TaqMan en casi 800 muestras del SNP candidato según los resultados obtenidos; en este caso, si se incluyó el sexo en el modelo ajustado además de las otras variables y se tomó como significativo un valor de $P < 0.05$.

Con la base de datos GWAS, se tomó como valor significativo una $P < 0.05$ corregida por Bonferroni (es decir, 0.05 / 719571 = P nominal < 6.95×10^{-8}) en los análisis de asociación básicos por SNP. Solo con el SNP más significativo en estos análisis, se corrieron regresiones logísticas ajustando por edad, sexo, ancestría global europea y ancestría global africana, con el fin de evaluar si la asociación se mantenía; en este caso, no se ajustó por la variable *"Array"*, pues todos los valores de ancestría usados fueron calculados con la misma base tipo GWAS. Adicionalmente, se realizaron análisis confirmatorios por TaqMan en casi 800 muestras. En los análisis que incluyeron solo el SNP candidato, se tomó como significativo un valor de $P < 0.05$.

Siguiendo las recomendaciones descritas en el protocolo de *Clarke GM, et al* (145), en todos los análisis se confirmaron las asociaciones sugestivas o significativas mediante la prueba de permutación de valores de *P* basado en 100 mil replicados de un modelo alélico ($P_{100\ mil\ permutaciones} < 0.05$). Los resultados de los análisis de asociación se resumieron por cromosoma en un gráfico tipo Manhattan, que permite ver los –log10(*P*) obtenidos para cada SNP. Se calculó el **factor de inflación lambda (λ)** para observar la correlación entre el –log10(*P*) observado versus el esperado en un gráfico quantil-quantil, y así descartar la influencia de factores confusores como la estratificación poblacional o sesgos de recolección o diferencias en el genotipado entre casos y controles de la muestra a estudio.

Con el fin de evaluar el papel de la variabilidad a nivel del genoma completo y de la ancestría local sobre el efecto de los SNPs encontrados como asociados, se corrieron modelos de regresión ajustados por edad, sexo, los 10 primeros componentes principales (PCs) que explican el 64% de la variabilidad genómica (ver **Figura - Anexos H**) y las proporciones de ancestría global, del cromosoma y del locus donde se ubica la variante a estudio; estos análisis solo fueron posibles en las muestras que tienen información tipo GWAS.

2.2.3 Resultados

Análisis de asociación de SNPs en genes candidatos (CG)

Para estos análisis CG se hicieron comparaciones por separado entre PA o CCR versus controles (ver **Tabla 19**). Al respecto, se encontraron diferencias significativas entre CCR versus controles a nivel de *P* nominal en la frecuencia del alelo menor (A) del SNP 14q11.2:rs1760898 (0.266 versus 0.381; OR 0.58; 95%CI 0.42-0.79; *P* nominal = 5.6×10^{-4}); sin embargo, no alcanzó niveles significativos al corregir por múltiples tests. Éste SNP se encontró en HWE en controles ($P > 0.05$).

Adicionalmente, se corrió una regresión logística ajustada de todos los SNPs incluidos en la base de datos CG (~1200) y los resultados sugieren un efecto protector del alelo menor del SNP 14q11.2:rs1760898 (A) con respecto al riesgo de CCR (OR 0.48; 95%CI 0.33-0.69; *P* nominal = 6.8 x 10^{-5}), pues se obtuvo una significancia marginal al corregir por múltiples test (*P* corregida por Bonferroni = 0.06) (ver **Tabla 20** y **Figura 10**B). No se evidenció estratificación poblacional (λ = 1.01) (ver **Figura 10**A). En base a que esta asociación sobrevivió el test de permutación ($P_{100\ mil\ permutaciones}$ = 0.03) (ver **Tabla 20**), se extendieron los análisis a muestras adicionales del estudio genotipadas por TaqMan, completando un n = ~ 800, y se corroboró el papel protector del alelo rs1760898 (A) con respecto al riesgo de CCR, tanto en los análisis básicos por SNP (*P* = 0.05) como en los análisis ajustados (OR 0.76; 95%CI 0.60-0.98*; P* = 0.03) (ver **Tabla 19** y **Tabla 20** respectivamente).

Por otro lado, en 404 muestras con información de genoma completo y resultados de TaqMan para el SNP rs1760898, se perdió la asociación con el riesgo de CCR en los análisis ajustados por la variabilidad genómica y las proporciones de ancestrías en el cromosoma 14, y en el locus 14q11.2, además de las otras variables (ver **Tabla 20**). No se observaron asociaciones entre el riesgo de PA y este SNP u otro incluido en la plataforma CG.

Tabla 19 Análisis básico de asociación por SNP *(X^2)* de la variante 14q11.2:rs1760898 *(TEP1)* con el riesgo de tumores colorrectales en colombianos

GENOTIPOS	N	FRECUENCIA DEL ALELO MENOR (A)		MODELO ALÉLICO (A VERSUS C)				
		AFECTADOS	NO AFECTADOS	OR	[95% CI]	*P* NOMINAL	*P* BONFERRONI	*P* 100 MIL PERMUTACIONES
CG (~ 1200 SNPs x 483 muestras)								
PA versus Controles	101 versus 218	0.307	0.381	0.72	[0.50-1.03]	0.07	1.00	1.00
CCR versus Controles	164 versus 218	0.262	0.381	0.58	[0.42-0.79]	**5.6×10^{-4}**	0.69	0.43
TaqMan (~ rs1760898 x 790 muestras)								
PA versus Controles	168 versus 339	0.307	0.347	0.83	[0.63-1.10]	0.20	-	0.24
CCR versus Controles	283 versus 339	0.293	0.347	0.78	[0.62-0.99]	**0.05**	-	**0.05**

Tabla 20 Regresiones logísticas ajustadas del SNP 14q11.2:rs1760898 *(TEP1)* con el riesgo de tumores colorrectales en colombianos

GENOTIPOS	MODELO ADITIVO (AA > AC > CC)								
	AJUSTADO POR EDAD + ANCESTRÍA GLOBAL EUROPEA + ANCESTRÍA GLOBAL AFRICANA					AJUSTADO POR EDAD + SEXO + 10 PCs + ANCESTRÍAS EUROPEA Y AFRICANA (GLOBALES, EN CROMOSOMA 14 Y EN LOCUS 14q11.2)α			
	OR	[95% CI]	*P* NOMINAL	*P* BONFERRONI	*P* 100 mil PERMUTACIONES	OR	[95% CI]	*P* NOMINAL	*P* 100 MIL PERMUTACIONES
CG (~ 1200 SNPs x 483 muestras)*									
PA versus Controles	0.73	[0.46-1.17]	0.19	1.00	1.00	-	-	-	-
CCR versus Controles¥	0.48	[0.33-0.69]	6.8×10^{-5}	0.06	**0.03**	-	-	-	-
CG (~ rs1760898 x 483 muestras)*									
PA versus Controles	0.73	[0.46-1.17]	0.19	-	0.19	-	-	-	-
CCR versus Controles	0.48	[0.33-0.69]	6.8×10^{-5}	-	**2.0×10^{-5}**	-	-	-	-
TaqMan (~ rs1760898 x 790 muestras)*§									
PA versus Controles	0.94	[0.70-1.28]	0.71	-	0.71	1.13	[0.74-1.74]	0.57	0.59
CCR versus Controles	0.76	[0.60-0.98]	**0.03**	-	**0.03**	1.30	[0.88-1.92]	0.19	0.21

* Adicionalmente ajustado por *"Array"*
§ Adicionalmente ajustado por sexo
α Análisis realizado solo con 404 muestras comunes entre las 790 y las 415 genotipadas con TaqMan y GWAS, respectivamente.
¥ Los resultados de éste análisis se resumieron por cromosoma en un gráfico tipo Manhattan, que permite ver los –log10(*P*) obtenidos para cada SNP (ver **Figura 10**B)

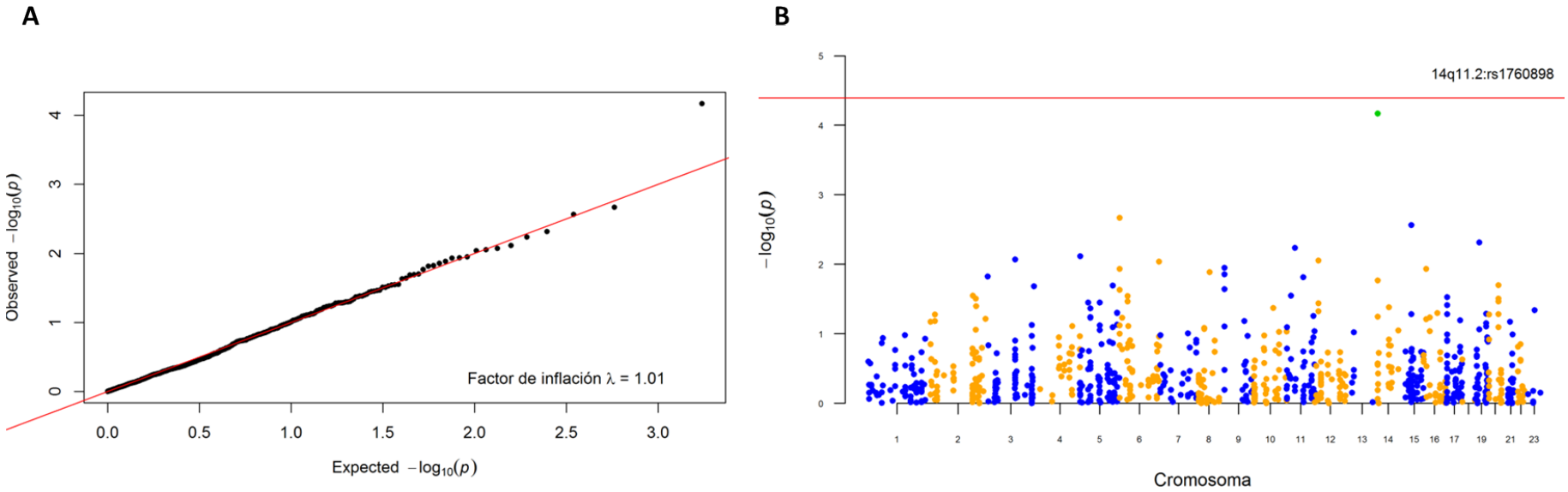


Figura 10 Resultados de los análisis de asociación de SNPs con CCR usando la aproximación CG en un modelo de regresión logística ajustado. A) Gráfico quantil-quantil de los –log10(*P*) observados y esperados. B) Gráfico tipo Manhattan de los –log10(*P*) de la asociación de SNPs con el riesgo de CCR, por cromosoma. El SNP más significativo se muestra en verde (14q11.2:rs1760898). Figura original generada en el programa *R statistics* (76).

Análisis de asociación de SNPs a nivel de genoma completo (GWAS)

Para estos análisis tipo GWAS, se hicieron comparaciones igualmente por separado entre PA o CCR versus controles (ver **Tabla 21**) y se encontraron diferencias significativas en la frecuencia del alelo menor (A) del SNP 17q25.3:rs1065768 al comparar PA con los controles (0.238 versus 0.473; OR 0.35, 95%CI 0.24-0.51; *P* nominal = 3.4 x 10^{-8}); esta asociación se mantuvo aún después de corregir por múltiples test (*P* corregida por Bonferroni = 0.02) (ver **Tabla 21** y **Figura 11**B). Como se observa en la **Figura 11**A, en estos análisis se evidenció poca estratificación poblacional importante (λ = 1.11). El SNP 17q25.3:rs1065768 se encontró en HWE en controles (*P* > 0.05).

Tabla 21 Análisis básico de asociación por SNP *(X^2)* de la variante 17q25.3:rs1065768 *(TK1)* con el riesgo de tumores colorrectales en colombianos

GENOTIPOS	N	FRECUENCIA DEL ALELO MENOR (A)		MODELO ALÉLICO (A VERSUS G)				
		AFECTADOS	NO AFECTADOS	OR	[95% CI]	*P* NOMINAL	*P* BONFERRONI	*P* 100 MIL PERMUTACIONES
GWAS (~ 700k SNPs x 415 muestras)								
PA versus Controles¥	122 versus 131	0.238	0.473	0.35	[0.24-0.51]	**3.4 x 10^{-8}**	**0.02**	**0.01**
CCR versus Controles	162 versus 131	0.413	0.473	0.78	[0.56-1.09]	0.14	1.00	1.00
TaqMan (~ rs1065768 x 783 muestras)								
PA versus Controles	165 versus 336	0.282	0.454	0.47	[0.36-0.63]	**1.7 x 10^{-7}**	-	**1.0 x 10^{-5}**
CCR versus Controles	282 versus 336	0.404	0.454	0.82	[0.65-1.02]	0.08	-	0.09

¥ Los resultados de éste análisis se resumieron por cromosoma en un gráfico tipo Manhattan, que permite ver los –log10(*P*) obtenidos para cada SNP (ver **Figura 11**B)

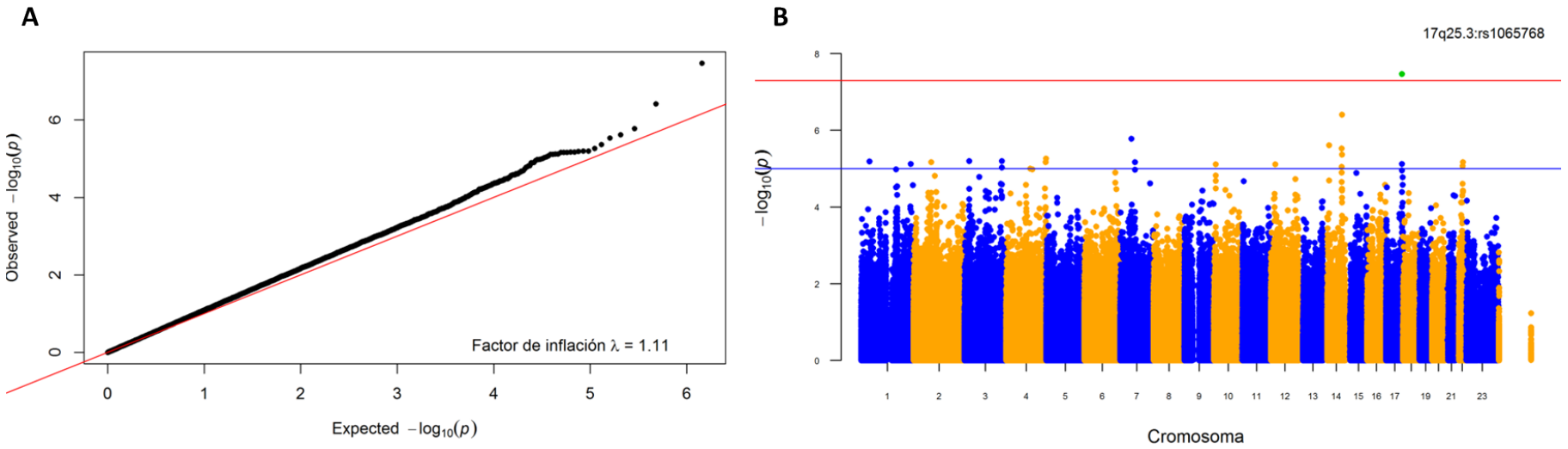


Figura 11 Resultados de los análisis de asociación básicos por SNP *(X^2)* con el riesgo de PA usando la aproximación GWAS. A) Gráfico quantil-quantil de los –log10(*P*) observados y esperados. B) Gráfico tipo Manhattan de los –log10(*P*) de la asociación de SNPs con el riesgo de PA, por cromosoma. El SNP más significativo se muestra en verde (17q25.3:rs1065768). Figura original generada en el programa *R statistics* (76).

Teniendo en cuenta que el análisis de regresión logística ajustada, usando todos los ~700k SNPs no convergió, éstos se corrieron solo seleccionando el SNP 17q25.3:rs1065768 y se obtuvieron resultados consistentes con un efecto protector del alelo menor (A) de este SNP con respecto al riesgo de PA (OR 0.38, 95%CI 0.24-0.58; *P* nominal = 8.4 x 10^{-6}) (ver **Tabla 22**). Más aún, dicha asociación se mantuvo cuando se incluyeron covariables adicionales como los 10 primeros PCs junto con las ancestrías europea y africana del cromosoma 17 y del locus 17q25 (OR 0.39, 95%CI 0.23-0.65; *P* nominal = 1.7 x 10^{-4}) (ver **Tabla 22**).

Estos resultados se replicaron en los análisis de las muestras genotipadas por TaqMan, completando un N de ~800, puesto que se mantuvo el efecto protector del alelo menor (A) del SNP 17q25.3:rs1065768 sobre el riesgo de PA, tanto en los análisis básicos por

SNP (P = 1.7 x 10^{-7}) como en los análisis ajustados (OR 0.54; 95%CI 0.40-0.74; P = 9.7 x 10^{-5}) (ver **Tabla 21** y **Tabla 22**, respectivamente). Por otro lado, se observó una tendencia del efecto protector de este SNP en el riesgo de CCR, con significancia marginal (OR 0.81; 95%CI 0.65-1.03; P = 0.08) (ver **Tabla 22**).

Tabla 22 Regresiones logísticas ajustadas del SNP 17q25.3:rs1065768 *(TK1)* con el riesgo de tumores colorrectales en colombianos

GENOTIPOS	MODELO ADITIVO (AA > AG > GG)								
	AJUSTADO POR EDAD + SEXO + ANCESTRÍA GLOBAL EUROPEA + ANCESTRÍA GLOBAL AFRICANA					AJUSTADO POR EDAD + SEXO + 10 PCs + ANCESTRÍAS EUROPEA Y AFRICANA (GLOBALES, EN CROMOSOMA 17 Y EN LOCUS 17q25.3)α			
	OR	[95% CI]	*P* NOMINAL	*P* BONFERRONI	*P* 100 mil PERMUTACIONES	OR	[95% CI]	*P* NOMINAL	*P* 100 MIL PERMUTACIONES
GWAS (~ 700k SNPs x 415 muestras)									
PA versus Controles	NA	NA	NA	NA	NA	-	-	-	
CCR versus Controles	NA	NA	NA	NA	NA	-	-	-	
GWAS (~ rs1065768 x 415 muestras)									
PA versus Controles	0.38	[0.24-0.58]	**8.4×10^{-6}**	-	**1.0×10^{-5}**	0.39	[0.23-0.65]	**1.7×10^{-4}**	**2.7×10^{-4}**
CCR versus Controles	0.77	[0.55-1.07]	0.12	-	0.12	0.79	[0.53-1.19]	0.26	0.28
TaqMan (~ rs1065768 x 783 muestras)*									
PA versus Controles	0.54	[0.40-0.74]	**9.7×10^{-5}**	-	**5.0×10^{-5}**	-	-	-	-
CCR versus Controles	0.81	[0.65-1.03]	0.08	-	0.08	-	-	-	-

* Adicionalmente ajustado por *"Array"*

α Análisis realizado solo con las 415 muestras genotipadas con GWAS

2.3 Discusión

Teniendo que cuenta que para el año que se terminaron de recolectar las muestras, habían reportado algunos SNPs asociados a CCR en poblaciones europeas, aprovechamos nuestras muestras colombianas para replicar estos hallazgos.

De los 14 SNPs previamente reportados, se encontró que el alelo menor del SNP 11q23:rs3802842 (C) se asoció como factor protector para CCR, según los análisis básicos de asociación por SNP *(X^2)*, pero que esta asociación fue marginal en los análisis ajustados por edad, sexo y ancestrías globales. A pesar que se usó una muestra de ~ 900 colombianos, es posible que la asociación no se haya mantenido por un bajo efecto de este SNP y falta de poder estadístico. Sin embargo, llama la atención que la dirección del efecto del alelo C de este SNP en colombianos es contrario a lo reportado en poblaciones europeas en las cuales este alelo aumenta el riesgo para CCR. Esto puede explicarse porque en poblaciones mezcladas como la nuestra, los bloques haplotípicos cambian debido a eventos de recombinación (162). Entonces, es posible que en nuestra población éste SNP se encuentre en LD con una combinación diferente de alelos en otros loci que pueden tener un efecto diferente en el riesgo CCR, en comparación con poblaciones más viejas como las europeas; lo anterior, especialmente teniendo en cuenta que se ha reportado que éste SNP no es el funcional sino que es un tagSNP de otra u otras variantes que sí parecen tener un efecto directo en el riesgo de tumores colorrectales (163).

Con respecto al alelo menor (A) del SNP 8q24:rs7014346, se observó una tendencia como factor de riesgo para CCR en colombianos y la dirección de ésta asociación está en acuerdo con lo reportado en poblaciones europeas. Al igual que en el caso anterior, la asociación solo fue marginalmente significativa. Ningún otro SNPs publicado se encontró asociado al riesgo de PA ni CCR en colombianos, en los modelos ajustados.

Una de las limitaciones de los estudios tipo GWAS en cáncer, es que las variantes de baja penetrancia asociadas al riesgo son difíciles de detectar; lo anterior, principalmente por el bajo tamaño del efecto de éstas o la baja frecuencia de los alelos de riesgo en la

población a estudio (51). Debido a esto, otras aproximaciones como la CG han tomado fuerza nuevamente (51). En este estudio se aplicaron dos aproximaciones, CG y GWAS, con el fin de identificar variantes nuevas de susceptibilidad para el riesgo de tumores colorrectales en Colombia, y se encontraron dos SNPs de protección para CCR y PA (14q11.2:rs1760898 y 17q25.3:rs1065768, respectivamente), no reportados antes.

Con respecto al SNP rs1760898 se encontró que el alelo menor (A) se asoció como factor protector para CCR, en los análisis de asociación tipo CG ($P_{100\ mil\ permutaciones} = 0.03$), lo cual fue confirmado por TaqMan. El SNP rs1760898 C>A está ubicado en el gen *TEP1* (Telomerase Associated Protein 1) y consiste en una **variante no sinónima** puesto que confiere un cambio de un aminoácido de tamaño mediano y polar (Asparagina, Asn) por otro de mayor tamaño y básico (Lisina, Lys), en la posición 307 de la proteína a lo largo del dominio TROVE, que consiste en un módulo de unión a RNA (UniProtKB/Swiss-Prot variant pages). La proteína TEP1 es un componente importante del complejo ribonucleoproteico conocido como telomerasa, responsable del elongamiento de los telómeros en los extremos de los cromosomas para proteger a la célula contra rearreglos genéticos grandes (161). Esta capacidad de elongar los telómeros disminuye en la mayoría de las células somáticas a medida que proliferan, lo que contribuye a su senescencia; por otro lado, en las células madre altamente proliferativas la longitud de los telómeros se mantiene gracias a la actividad telomerasa (161).

Teniendo en cuenta lo anterior, es posible que por las características físico-químicas ligadas al cambio de aminoácido alelo-específico del SNP rs1760898 C>A (p.Asn307Lys), se genere una interferencia con la unión al RNA que es necesario para el elongamiento telomérico. Lo anterior, puede indicar que individuos con el alelo de protección en este locus tienen telómeros más cortos y por lo tanto, no adquieren fácilmente el fenotipo hiperproliferativo que es característico de las células cancerosas, explicando así el efecto protector de la variante A. Si bien se requieren estudios adicionales para confirmar esta hipótesis, ésta puede estar soportada por otros que encontraron que el alelo mayor (C) del SNP 3q26.2:rs10936599 en el gen *hTERC*, asociado a un mayor riesgo de CCR y PA en poblaciones del Reino Unido (38, 152) y Asia (55), también se asoció a un aumento en la longitud de los telómeros en linfocitos periféricos de pacientes con CCR; por lo que concluyeron que los telómeros largos son

un factor de riesgo para CCR al simular un fenotipo de células madre altamente proliferativas (161).

Con respecto al SNP rs1065768 se encontró que el alelo menor (A) se asoció como factor protector para PA en los análisis básicos por SNP de los datos tipo GWAS (*P* nominal = 3.4 x 10^{-8}, *P* corregido por Bonferroni = 0.02) y ésta asociación fue consistente en los análisis confirmatorios por TaqMan realizados en ~ 800 muestras. El SNP rs1065768 G>A está ubicado en la región regulatoria 3'UTR del gen *TK1* (Thymidine kinase 1), que codifica una proteína involucrada en la síntesis del DNA y proliferación celular (UniProtKB/Swiss-Prot variant pages). Específicamente, la proteína TK1 es una quinasa que cataliza la conversión de desoxitimidina a desoxitimidina monofosfato, mediante la adición de un grupo fosfato a partir de una molécula de ATP; paso indispensable en la síntesis del DNA. De acuerdo a una búsqueda en miRBase (www.mirbase.org), en presencia del alelo A, el miRNA "hsa-miR-365b-5p" podría unirse debido a su homología con esta secuencia, a diferencia del alelo G. Esto puede indicar que el alelo A favorece la regulación postranscripcional del mRNA de *TK1*, afectando directamente la proliferación celular. Lo anterior, teniendo en cuenta que existen reportes de que los SNPs pueden influenciar la afinidad de unión con miRNAs al generar o destruir sitios de unión en las moléculas de mRNAs blanco (164). Nuevamente, se requieren estudios adicionales para probar el papel de este alelo en la regulación del ciclo celular y la tumorogénesis; especialmente, debido a que encontramos evidencia del papel protector del alelo menor (A) en PA y una tendencia en CCR.

La asociación más consistente se observó para la variante rs1065768 con el riesgo de PA, pues se mantuvo aún después de corregir por los 10 PCs y las ancestrías globales, en cromosoma 17 y en el locus 17q25.3, en una muestra de 415 colombianos. Por otro lado, la asociación del SNP rs1760898 con el riesgo de CCR no se mantuvo al corregir por la variabilidad a nivel de genoma completo y las ancestrías globales, en cromosoma 14 y en el locus 14q11.2, en una muestra de 404 colombianos; lo anterior puede deberse al bajo efecto de esta variante en el riesgo de CCR que junto con el bajo tamaño de la muestra usado para estos análisis complejos, no se alcanzó el poder estadístico suficiente para mantener dicha asociación. Otra posible razón es que se trate de un tagSNPs de otra variante con un mayor efecto en la enfermedad, que no fue posible detectar por la baja densidad de SNPs incluidos en la plataforma CG.

La variante rs1760898 no se asoció con el riesgo de PA; mientras que si se vio una tendencia en la asociación de la variante rs1065768 con el riesgo de CCR. Lo anterior puede deberse a nuestro moderado tamaño de muestra o a que solo el 10% de los adenomas progresan a cáncer (153), luego es de esperarse que existan variantes comunes y variantes únicas para cada caso. Más aún, solo algunas de las variantes comunes de riesgo para CCR, hasta ahora reportadas en estudios tipo GWAS, se han evidenciado como de riesgo para PA.

Se puede concluir de éste capítulo que se lograron identificar dos nuevas variantes genéticas de riesgo en los colombianos, no antes reportadas en otras poblaciones. Las dos asociaciones se mantuvieron aun después de corregir por posible estratificación poblacional, en una muestra multiregional de Colombia con alto grado de mestizaje. Estos resultados son de gran aporte no solo para avanzar en el entendimiento de las bases biológicas del desarrollo tumoral en colon y recto, sino también para explicar parte de los casos de CCR en nuestro país, contribuyendo así, al conocimiento general del papel de la heredabilidad en el riesgo de ésta enfermedad y lesiones premalignas asociadas. Más aun, éstos SNPs merecen ser validados en una muestra independiente para evaluar su potencial como marcadores de susceptibilidad y su uso en la estratificación de colombianos con un mayor riesgo de tumores colorrectales; la identificación de éstos individuos, podría servir de filtro para el ofrecimiento de opciones efectivas de tamizaje para el diagnóstico temprano de la enfermedad, lo cual tendría un alto impacto a nivel de salud pública en nuestro país.

3. Perfil proteómico de tumores colorrectales en colombianos

Aspectos teóricos relacionados con el estudio de perfiles proteómicos y sus métodos de abordaje:

De acuerdo al *Dogma Central de la Biología Molecular* descrita por Francis Crick en 1958, el flujo de la información genética procede de la transcripción del DNA a RNA y de la traducción del RNA a PROTEÍNA (165). El conjunto de proteínas expresado por un organismo, tejido o célula en un estado biológico particular se conoce como *proteoma* y la disciplina que lo estudia se denomina *proteómica* (166-168). Teniendo en cuenta que el proteoma en un espécimen y en un momento determinado es único y específico, se considera que las estrategias en proteómica para el estudio de enfermedades como el cáncer, son herramientas innovadoras, dinámicas y complementarias a otras metodologías a gran escala como la genómica y la transcriptómica, entre otros (167). Más aun, el estudio de los perfiles proteómicos en bioespecímenes obtenidos de pacientes con cáncer podrían aportar a nuestro conocimiento en la patogénesis del tumor y al descubrimiento de marcadores tempranos de enfermedad o de progresión, así como de nuevos blancos terapéuticos (167).

Los estudios en proteómica se basan en la separación de las proteínas contenidas en una muestra biológica mediante técnicas de alta resolución (169), como son la electroforesis bidimensional en geles de poliacrilamida, 2DE (170) o la cromatografía liquida de alta eficiencia, HPLC (171), ambos acoplados a la identificación de proteínas mediante espectrometría de masas en tándem, MS-MS (172). Sin embargo, independientemente del método de separación empleado, es importante tener en cuenta los factores relacionados con la toma, almacenamiento y procesamiento de las muestras

a analizar, pues todos estos influyen en la calidad de los proteomas separados y por consiguiente, en la confiabilidad de los resultados (173).

A continuación, se describen los fundamentos generales de los métodos o abordajes a gran escala más usados en proteómica:

i) La 2DE es un método ortogonal para separar mezclas complejas de proteínas basándose en dos de sus propiedades, el **punto isoeléctrico (pI)** en una primera dimensión y la masa molecular en la segunda dimensión (174). La primera separación se conoce como **isoelectroenfoque** y se basa en el uso de una tira de gel de acrilamida con un **gradiente de pH inmóvil** (tiras IPG) sobre la cual se aplica un potencial eléctrico; esta combinación permite separar las proteínas de una mezcla a lo largo de la tira, hasta alcanzar el pH donde su carga neta es cero (174). Una vez llevados a cabo los pasos de equilibrio que incluyen la reducción de puentes disulfuro y alquilación o bloqueo de los grupos tiol de las proteínas separadas en la tira, esta se acopla a la segunda dimensión para continuar con la separación de las proteínas por masa molecular por medio de una electroforesis en gel de poliacrilamida en condiciones denaturantes (SDS-PAGE) (169). Con esta metodología es posible identificar spots diferenciales en los mapas proteómicos entre diversas condiciones a comparar, con el fin de seleccionarlos para posteriores estudios dirigidos a la identificación de las proteínas diferencialmente expresadas entre los grupos (169, 175). Sin embargo, factores relacionados con la reproducibilidad de la técnica 2DE, debido a múltiples pasos de manipulación de la muestra y la falta de robustez y reproducibilidad de los métodos usados en el proceso de comparación de la intensidad de los spots entre geles, generan limitaciones posteriores en los pasos de la identificación y cuantificación de las proteínas mediante MS-MS (175).

ii) Los métodos en HPLC son un tipo de cromatografía de elución que consiste en el paso de la muestra a separar dentro de una fase móvil (líquida) a través de una fase estacionaria (líquido inmiscible o sólido), en donde cada analito avanzará a lo largo del sistema con una velocidad diferente que dependerá de la afinidad del mismo con cada una de las dos fases, por lo cual eluirán al final del sistema en tiempos diferentes (175, 176). Una característica muy

importante de este método, es su versatilidad, pues permite el uso de diferentes fases móviles y estacionarias para la separación de muestras complejas basándose en diversas propiedades fisicoquímicas, mejorando así la capacidad de separación según el interés (176). La cromatografía líquida en fase reversa (RPLC) es uno de los métodos más usados, mediante el cual es posible separar una mezcla compleja según la **hidrofobicidad** de las moléculas (175, 176). Este sistema se caracteriza por el uso de una fase estacionaria apolar, por ejemplo una columna C18 que consiste en una columna de silica con ligandos hidrofóbicos inmovilizados de 18 carbonos, que en presencia de una fase móvil polar permitirá la retención diferencial de las moléculas apolares por la columna y la elución de las moléculas polares con la fase móvil (175, 176). El tiempo de retención de las moléculas va cambiando según su hidrofobicidad y su afinidad con las fases durante la aplicación de la fase móvil en gradiente; es decir, las moléculas apolares inicialmente retenidas por la columna en presencia de una fase móvil polar, se van eluyendo del sistema a medida que se añade un mayor porcentaje de solventes hidrofóbicos o menos polares en la fase móvil (175, 176). Por convención, el solvente A es el acuoso o polar, como es por ejemplo el agua grado HPLC, y el solvente B es un solvente orgánico de polaridad media miscible en agua, como el acetonitrilo, el metanol o el propanol (176). Cuando la muestra a analizar contiene grupos funcionales ácidos o básicos (ionizables), es importante controlar el pH dentro del sistema añadiendo bajos porcentajes de soluciones ácidas como el ácido fórmico, acético o trifluoroacético a la fase móvil polar (solvente A), con el fin de proveer una fuente adicional de protones y de esta manera suprimir los fenómenos de ionización que pueden cambiar la polaridad de los compuestos; el control de estos eventos permitirá tener una menor variabilidad en los tiempos de retención durante la separación entre las fases y mejorar la reproducibilidad de la corrida cromatográfica (176). Una ventaja importante de este método, es que se alcanza una alta resolución en la separación de los analitos, disminuyendo la no separación o co-elución de éstos y mejorando su identificación en el espectrómetro de masas, MS (175, 176). Otra gran ventaja del HPLC en el estudio de perfiles proteómicos, es que permite separar una gran diversidad de compuestos con diferente polaridad, termo lábiles/estables,

volátiles/no volátiles y de alto/bajo peso molecular (177). Se requiere de una interfase para el acople entre el sistema de separación HPLC (estado líquido y de alta presión) con el sistema de identificación y cuantificación MS (estado gaseoso y en condiciones de vacío), como se puede ver en la **Figura 12** (177).

iii) La MS-MS ha permitido avanzar en los métodos de cuantificación relativa de la expresión de proteínas, a partir del cual se obtiene una medida de los cambios en la abundancia de las proteínas de interés en un grupo de estudio con respecto al otro, expresado como *fold change* (178). En la actualidad existen muchas tecnologías MS disponibles, las cuales ofrecen diferentes niveles en cuanto a sensibilidad y resolución en la identificación y cuantificación de los componentes de una mezcla. Para lograr lo anterior, el equipo MS debe estar conformado por: i) una interfase que consta con una fuente de ionización y un **sistema de desolvatación** usados para la conversión de los analitos de una muestra en iones; ii) un analizador de masas en donde los iones son separados en función de su relación masa/carga (m/z); y iii) un detector que recibe y convierte el haz de iones de determinada m/z en una señal eléctrica, la cual es posteriormente procesada para generar el espectro de masas (ver **Figura 12**) (177, 179, 180). Todas las etapas anteriores se realizan en condiciones de estricto vacío (~ 10^{-5} a 10^{-8} torr), con el objetivo de bajar al máximo la presión del sistema y guiar a los iones hacia el detector sin que choquen o se dispersen (179). Dentro de las interfases más recomendadas en el estudio proteómico de muestras complejas, se encuentran los que usan el método de *ionización por electrospray, ESI*; esto se debe a que logra la ionización de biomoléculas con un amplio rango de polaridad (180). Adicionalmente, al ser un método de ionización en líquido que no requiere una previa volatilización de los compuestos, ESI permite la ionización de analitos volátiles, no volátiles, termolábiles y termoestables, favoreciendo de esta manera el análisis de una mayor cantidad de compuestos en comparación con otras interfases (177, 181).

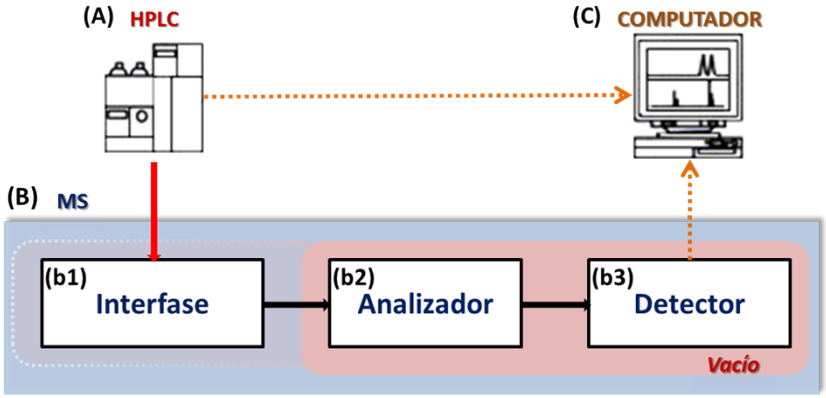


Figura 12 Esquema de los componentes de un sistema HPLC-MS. A) Equipo *HPLC*, que maneja flujo líquido a un alto volumen y presión (fase móvil). B) Equipo de MS, conformado por: b1) La interfase, lugar donde ocurre la transición de los analitos en solución a iones, los cuales se forman por: i) ionización gaseosa o ii) ionización en líquido; lo anterior, en condiciones de vacío o a presión atmosférica según el tipo de interfase y antes de ser dirigidos al analizador. b2) El analizador, donde se separan los iones según su relación m/z y son guiados al detector gracias a la aplicación de diferentes potenciales de radiofrecuencia (Rf), campos eléctricos o magnéticos, en condiciones de alto vacío. b3) El detector, donde se reciben los haz de iones o grupos de iones (m/z). C) Computador, donde se recibe y se procesa la información que llega del detector para generar el espectro de masas, el cual consiste en un gráfico de barras. Cada barra representa un ión con determinada relación m/z (eje X) y el alto de la barra corresponde a la abundancia relativa del ión (intensidad relativa, %, en el eje Y); esta abundancia relativa está dada en relación al pico base (ión más frecuente al cual se le asigna una abundancia del 100%). Figura modificada de *http://www.shimadzu.com/an/lcms/support/intro/lib/lctalk/46/46intro.html*

Con respecto a los analizadores, también existen varias opciones con diferente capacidad de resolución y precisión en la identificación y cuantificación de proteínas en una mezcla compleja. Por su alta resolución, precisión y sensibilidad, además del bajo costo y baja complejidad instrumental, los MS que incluyen analizadores de tipo *Orbitrap* son de los más recomendados en los estudios proteómicos; lo anterior, incluso en comparación con otros de alta resolución como el analizador *FTICR*, el cual se caracteriza por ser mucho más costoso, instrumentalmente masivo y tener una baja resolución en la identificación de compuestos de alto peso molecular (182, 183). El acople HPLC/MS-MS, usando ESI + Q-Exactive Orbitrap como sistema MS, ofrece múltiples ventajas sobre otros sistemas de acople; pues permite ionizar una gran cantidad de analitos que pueden ser posteriormente filtrados o seleccionados según su relación m/z en un cuadrupolo, desde donde son guiados de manera pulsada a una trampa de iones en forma de C, (C-trap) (183-185). Estos pulsos o paquetes de iones están determinados por

el parámetro AGC (Automatic Gain Control) mediante el cual se limita la entrada de una población de iones en la trampa, con el fin de eliminar o minimizar los efectos de carga o de repulsión entre las partículas de igual carga dentro de la trampa, mejorando así la precisión en la estimación de las masas de interés (183-186). Por último, el acople del C-trap con una celda de **disociación por colisión de alta energía, HCD**, y con el analizador Orbitrap, permite realizar múltiples ciclos de fragmentación y detección de los iones seleccionados de forma rápida y precisa, sin necesidad de un detector adicional (186, 187).

Antecedentes de los métodos de tamizaje, de diagnóstico y de pronóstico en CCR y avances en la identificación de potenciales biomarcadores mediante estudios proteómicos:

Actualmente, la mayoría de países usan la *sangre oculta en materia fecal* como prueba de tamización del CCR; si bien es una prueba barata, no invasiva, de fácil implementación y con la cual se ha observado un efecto en el aumento de la supervivencia, tiene la desventaja de generar altas tasas de falsos positivos y negativos por la baja sensibilidad y especificidad como prueba diagnóstica (188). Por otro lado, la *colonoscopia* permite la detección de lesiones preneoplásicas y de neoplasias aún en estadios tempranos de la enfermedad, por lo que su implementación tiene un impacto muy importante en la disminución de las tasas de incidencia y mortalidad por ésta enfermedad (188). Acompañada de la biopsia, esta técnica logra una alta sensibilidad (97%) y especificidad (98%) en el diagnóstico del CCR, por lo que se considera actualmente como el *gold standard* (188). Sin embargo, es operador-dependiente, invasiva, costosa, requiere de una preparación especial y es en general un procedimiento incómodo para los pacientes que trae consigo riesgos para la salud (188). Adicionalmente, estudios sugieren que la colonoscopia no detecta entre el 2% al 6% de los tumores de ~ 1cm, lo que reduce su efectividad en la prevención y diagnóstico temprano de la enfermedad (189, 190).

Los biomarcadores consisten en biomoléculas que reflejan cambios bioquímicos, fisiológicos o morfológicos, son medibles y se asocian a fenotipos o condiciones

específicas; además, estos pueden ser medidos directamente en el tejido de interés, o incluso en otros más accesibles y menos invasivos como son la sangre y orina (191-193). Los cambios moleculares implicados en el desarrollo de CCR toman varios años (13) y algunos biomarcadores podrían ser empleados para detectar la enfermedad en estadios tempranos, lo que es esencial para reducir las tasas de mortalidad (194).

Dentro del proceso de validación de posibles biomarcadores se busca que sean altamente sensibles y específicos para el diagnóstico de la enfermedad de interés, o para fines pronósticos; sin embargo, gran parte del éxito en su implementación en la práctica clínica, está en lograr su medición con alta precisión y robustez en fluidos de fácil acceso, garantizando la aceptación y comodidad de los pacientes (194).

Por el momento, el único biomarcador en sangre aprobado por la FDA es el antígeno carcino-embrionario, CEA, el cual se usa como biomarcador de pronóstico o recurrencia de la enfermedad con una sensibilidad que aumenta con el estadio clínico; sin embargo, no es específico de CCR pues puede estar aumentado en enfermedad inflamatoria intestinal, pancreatitis, alteraciones hepáticas y otras neoplasias (188, 194). Otro biomarcador de pronóstico en sangre para el CCR, es el antígeno carbohidrato 19-9, CA19-9, cuyo uso es más limitado al ser menos sensible y específico que el CEA (188, 194).

En respuesta a la carencia de biomarcadores altamente sensibles y específicos en CCR, se ha avanzado en la implementación de estudios proteómicos en diversos tipos de muestras biológicas para el análisis en simultáneo de la mayor cantidad de proteínas posibles, y se han logrado proponer patrones o perfiles de expresión diferenciales; sin embargo, los candidatos identificados solo se han verificado en muestras pequeñas independientes (195, 196). Es decir que ninguno de éstos potenciales biomarcadores están en proceso de validación en estudios de cohorte con adecuados tamaños de muestra que permitan probar su utilidad en la práctica clínica; esto se debe en parte a la dificultad en conseguir las muestras para este fin, especialmente si se trata de tejido colorrectal (195).

El descubrimiento de marcadores mediante estudios proteómicos en muestras biológicas de fácil acceso, como derivados sanguíneos, podría facilitar la conducción de posteriores

estudios de validación a nivel poblacional, puesto que este tipo de muestras pueden captarse a través de biobancos (197). Al respecto, es preciso tener en mente que con este abordaje se enfrentarán otros retos, los cuales están relacionados con el amplio rango dinámico en la concentración de las proteínas en suero y plasma que abarca más de 10 órdenes de magnitud; esto disminuye el número de proteínas identificables debido a que proteínas muy abundantes, como la albúmina e inmunoglobulinas, enmascaran otras menos abundantes (198).

3.1 Estudio exploratorio de perfiles proteómicos de pacientes colombianos con CCR

En la introducción a este capítulo, se expuso la necesidad que existe de encontrar biomarcadores en CCR que tengan potencial como herramientas para el tamizaje, diagnóstico o pronóstico de la enfermedad, con una alta sensibilidad y especificidad, y que además sean mínimamente invasivos y accesibles. Teniendo en cuenta esta necesidad, se planteó como ***propósito de esta quinta parte del estudio, explorar diferencias en los perfiles proteómicos en plasma entre pacientes con CCR y controles.***

3.1.1 Objetivos específicos

3.1.1.1 Evaluar diferencias en la expresión de proteínas en plasma entre pacientes con CCR y controles del estudio.

3.1.1.2 Proponer proteínas que actúen como posibles biomarcadores de diagnóstico o pronóstico para CCR

3.1.2 Métodos

Para este estudio de perfiles proteómicos del CCR, se incluyeron muestras que habían sido captadas para el estudio de marcadores de susceptibilidad genética, expuesto en los capítulos anteriores. En total, se seleccionaron muestras de plasma de 20 casos de CCR y 6 controles, teniendo en cuenta los siguientes criterios: i) sexo masculino; ii) edad entre 45 a 65 años; iii) que la muestra de sangre se hubiera tomado en ayuno; y iv) que se

hubiera autorizado el uso de las muestras en estudios futuros. *Ver el Anexo 2 – Materiales y métodos en análisis proteómicos, para obtener información más detallada de los métodos.*

Teniendo en cuenta el carácter exploratorio de este estudio, se prefirió implementar la modalidad *"shutgun proteomics"* con la cual se busca analizar mezclas complejas de proteínas (199); diferente a la modalidad de *"targeted proteomics"* la cual se centra en el análisis de proteínas específicas (200). Dentro de la modalidad seleccionada, se escogió el modo *"bottom-up"*, que hace referencia a la caracterización de las proteínas a partir del análisis de péptidos obtenidos de la digestión controlada de proteínas intactas; es decir, se parte del análisis de fragmentos peptídicos para inferir información sobre las proteínas en una mezcla (199).

La preparación de las muestras de plasma incluyó: i) obtención de plasma libre de plaquetas y ii) reducción de la complejidad mediante la inmunodepleción de albúmina e inmunoglobulina G[4]. Los péptidos a analizar se obtuvieron mediante digestión en gel con tripsina. Este método tiene como ventaja la remoción de contaminantes durante la electroforesis, como sales y detergentes, por lo que permite una alta compatibilidad con los análisis MS (199).

Los péptidos se separaron mediante RPLC, usando una columna C18 como fase estacionaria y una fase móvil compuesta por 0.1% de ácido fórmico, FA, (solvente A) y un gradiente de acetonitrilo, ACN, (solvente B). Los péptidos eluidos fueron ionizados en una interfase tipo ESI y posteriormente separados, fragmentados y detectados en un equipo Q-Exactive Orbitrap, usando el modo dependiente de datos con un SCAN MS de iones precursores, seguido de 15 SCANs MS-MS. Los espectros de masas se analizaron con el programa X! Tandem, usando como referencia la base de datos Uniprot Human Reference y se aceptaron identificaciones de péptidos con una probabilidad ≥ 99% (False Discovery Rate, FDR de 0.6%) e identificaciones de proteínas con una probabilidad ≥

[4] Para más información con respecto al método usado para la reducción de la complejidad del plasma, remitirse al artículo original publicado: *Rodríguez RA, Urrego WA, Sanabria-Salas MC, et al. Implementación de una metodología para la separación de proteomas de plasma humano mediante electroforesis bidimensional. Revista Colombiana de Química. 2015;44(3):30-38*

95%, con mínimo 2 péptidos (FDR de 0.02%) usando el programa Scaffold. Con estos criterios, se obtuvo un total de 469 proteínas y 437,438 péptidos.

La cuantificación relativa de las proteínas y los análisis de perfiles proteómicos diferenciales se realizaron usando el conteo espectral total o cuantificación sin marcaje, los cuales se normalizaron de acuerdo a los requerimientos de los paquetes y funciones usadas en R statistics (76). Se realizó un **análisis de componentes principales, PCA**, usando los dos primeros PCs de la variabilidad en la expresión de proteínas, con el fin de probar su capacidad para separar los dos grupos (casos y controles). Se corrieron regresiones lineales generalizadas, GLM, para encontrar evidencia en la expresión diferencial de proteínas entre los dos grupos, y se seleccionaron las más significativas de acuerdo a un valor de *P* corregido por FDR < 0.05 y un *fold change* ≥ 2.0 (el doble de expresión) o ≤ 0.5 (expresión a la mitad). Con las proteínas seleccionadas se realizó un análisis de clasificación jerárquica no supervisado tipo HeatMap, y los resultados se confirmaron mediante otro método probabilístico de clusterización no supervisado.

3.1.3 Resultados

Expresión diferencial de proteínas en plasma entre pacientes con CCR y controles del estudio.

En la **Tabla 23** se muestran las características generales de los individuos cuyas muestras de plasma fueron seleccionadas para los análisis proteómicos. El 100% de los individuos correspondieron al sexo masculino. La media de la edad en los controles fue de ~ 58 años (n = 6; 95%CI 51.8-63.9), mientras que en los casos fue de ~ 55 años (n = 20; 95%CI 52.7-58.1). El 75% de los casos de CCR fueron de localización izquierda y en su gran mayoría correspondieron a cáncer de recto; solo el 25% de los casos incluidos eran in situ al momento del diagnóstico. El 60% de los casos de CCR y el 50% de los controles, eran originarios de la región costera del país.

Tabla 23 Características de casos y controles seleccionados para el análisis proteómico

CARACTERÍSTICAS	CONTROLES n = 6	CÁNCER COLORRECTAL (CCR) n = 20
Edad		
Mínimo	50.00	45.00
1er cuartil	53.50	50.75
Mediana	59.00	56.00
Media [95% CI]	57.83 [51.81-63.86]	55.40 [52.67-58.13]
3er cuartil	62.25	59.75
Máximo	64.00	65.00
Diagnóstico		
Cáncer de colon	NA	6
Cáncer de recto	NA	14
Grado de invasión		
In situ	NA	5
Invasivo	NA	15
Localización		
Derecho	NA	5
Izquierdo	NA	15
Región		
Andina	3	8
Costera	3	12

La **Figura 13** muestra los resultados del PCA. Como se puede observar, los dos primeros PCs explican el 25% y el 16% de la varianza de los datos de expresión proteica, con los cuales fue posible la separación de la mayoría de las muestras según su fenotipo de cáncer o control. Sin embargo, dos muestras de individuos controles (con códigos 68287 y 8023) muestraron mayor semejanza con los casos. Igualmente, se observa una variabilidad biológica importante, especialmente entre los individuos con cáncer.

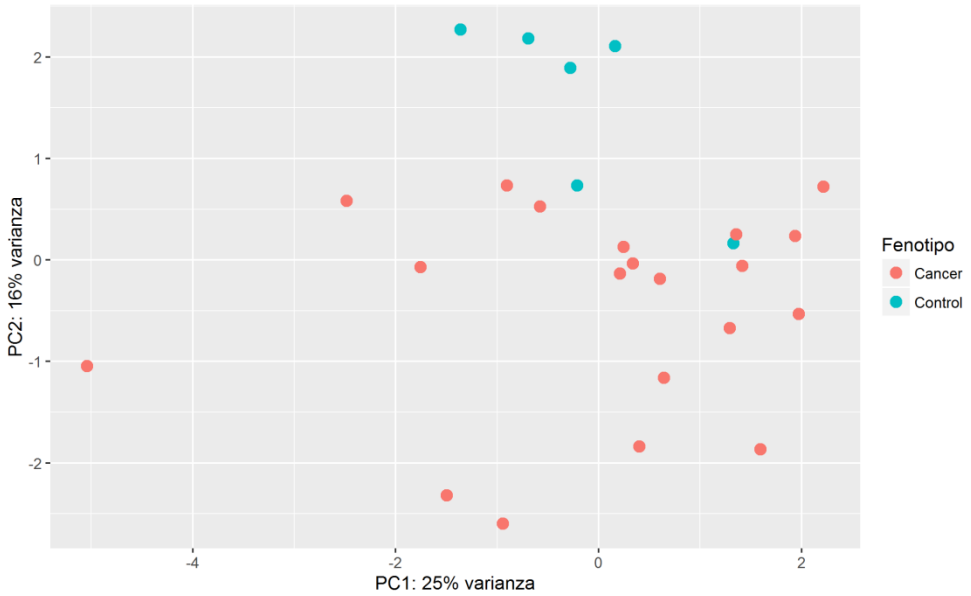


Figura 13 PCA de las muestras incluidas en el análisis proteómico. Figura original generada en el programa *R statistics* (76).
PC1, componente principal 1; PC2, componente principal 2

En las **Tabla 24**, **Tabla 25** y **Tabla 26** se listan las proteínas que mostraron diferencias a nivel de *P* no corregida < 0.05 en los análisis de expresión diferencial realizados solo en los casos de CCR, según el diagnóstico, el grado de invasión y la localización del cáncer, respectivamente; como se puede observar, ninguna de éstas sobrevivió la corrección por múltiples tests (FDR > 0.05) y tampoco mostraron diferencias importantes en el *fold change*, según los criterios establecidos.

Tabla 24 Proteínas expresadas diferencialmente entre casos de cáncer de recto comparado con casos de cáncer de colon

UNIPROT	ANÁLISIS DE EXPRESIÓN DIFERENCIAL EN SOLO CASOS DE CCR (Proteínas ~ Diagnóstico)				
	Media *	log2(Fold Change)	Fold Change	*P*	FDR
J3QLC9_HUMAN	474.90	0.388	1.3	0.05	1.00
IGHA1_HUMAN	268.72	-0.407	0.8	0.04	1.00
IGKC_HUMAN	94.01	-0.455	0.7	0.02	1.00
C1R_HUMAN	18.62	0.346	1.3	0.05	1.00
C1S_HUMAN	17.74	0.455	1.4	0.02	1.00
KV106_HUMAN	15.48	-0.513	0.7	0.01	1.00
F13A_HUMAN	3.99	-0.657	0.6	0.03	1.00
LV105_HUMAN	2.35	-0.618	0.7	0.04	1.00
PIGR_HUMAN	1.12	-0.511	0.7	0.04	1.00
LV102_HUMAN	1.15	-0.624	0.6	0.03	1.00

Los valores de *P* y FDR corresponden al análisis GLM realizado para evaluar diferencias en la expresión de proteínas entre grupos, y la medida de las diferencias de expresión entre grupos es reportada como log2(*Fold Change*); este valor fue posteriormente usado para calcular el *fold change*. Las proteínas están listadas en orden descendente de acuerdo a la Media (FDR = 1.00).
* Media de la expresión de cada proteína en el grupo de casos de cáncer de colon (referencia)

Tabla 25 Proteínas expresadas diferencialmente entre casos de CCR in situ comparado con casos de CCR invasivo

UNIPROT	ANÁLISIS DE EXPRESIÓN DIFERENCIAL EN SOLO CASOS DE CCR (Proteínas ~ Grado de invasión)				
	Media *	log2(Fold Change)	Fold Change	*P*	FDR
ITIH3_HUMAN	17.23	0.710	1.6	2.2×10^{-3}	0.30
B0YIW2_HUMAN	5.40	-0.858	0.6	2.4×10^{-3}	0.30
A2MG_HUMAN	804.44	-0.379	0.8	0.05	1.00
CFAB_HUMAN	76.66	0.271	1.2	0.02	1.00
AMBP_HUMAN	35.42	-0.346	0.8	0.03	1.00
APOE_HUMAN	25.45	-0.405	0.8	0.04	1.00
Q5VY30_HUMAN	23.49	-0.667	0.6	0.01	1.00
APOD_HUMAN	21.46	-0.393	0.8	0.05	1.00
PP2BB_HUMAN	0.93	0.561	1.5	0.03	1.00

Los valores de *P* y FDR corresponden al análisis GLM realizado para evaluar diferencias en la expresión de proteínas entre grupos, y la medida de las diferencias de expresión entre grupos es reportada como log2(*Fold Change*); este valor fue posteriormente usado para calcular el *fold change*. Las proteínas están listadas en orden ascendente de acuerdo al FDR.
* Media de la expresión de cada proteína en el grupo de casos de CCR in situ (referencia)

Tabla 26 Proteínas expresadas diferencialmente entre casos de CCR derecho comparado con casos de CCR izquierdo

UNIPROT	ANÁLISIS DE EXPRESIÓN DIFERENCIAL EN SOLO CASOS DE CCR (Proteínas ~ Localización)				
	Media *	**log2(Fold Change)**	**Fold Change**	***P***	**FDR**
KV106_HUMAN	15.47	-0.609	0.7	2.2×10^{-3}	0.55
ITIH1_HUMAN	81.32	0.431	1.3	4.5×10^{-3}	0.55
IGKC_HUMAN	94.01	-0.544	0.7	0.01	0.56
LG3BP_HUMAN	3.96	-0.744	0.6	0.02	0.99
ANT3_HUMAN	80.42	0.282	1.2	0.03	1.00
KV305_HUMAN	12.80	-0.504	0.7	0.03	1.00
LV105_HUMAN	2.35	-0.610	0.7	0.05	1.00
PIGR_HUMAN	1.12	-0.568	0.7	0.03	1.00
LV205_HUMAN	1.16	-0.644	0.6	0.03	1.00

Los valores de *P* y FDR corresponden al análisis GLM realizado para evaluar diferencias en la expresión de proteínas entre grupos, y la medida de las diferencias de expresión entre grupos es reportada como log2(*Fold Change*); este valor fue posteriormente usado para calcular el *fold change*. Las proteínas están listadas en orden ascendente de acuerdo al FDR.
* Media de la expresión de cada proteína en el grupo de casos de CCR derecho (referencia)

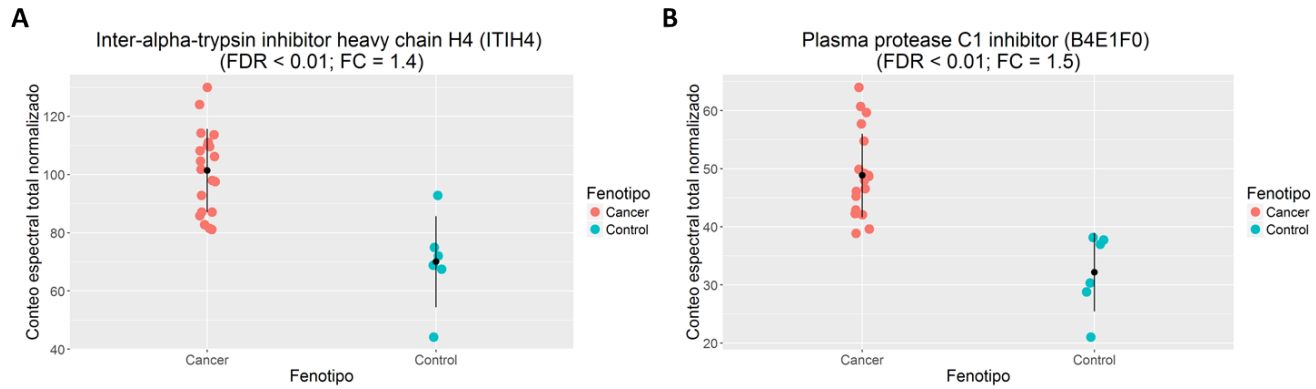
Teniendo en cuenta que ninguna de las comparaciones anteriores reveló resultados con suficiente significancia estadística, se decidió agrupar todos los casos de CCR y compararlos con los controles, con el objetivo de obtener diferencias más consistentes. Como se puede observar en la **Tabla 27**, estos análisis permitieron evidenciar diferencias significativas en la expresión de 14 proteínas entre los dos grupos con un FDR < 0.05.

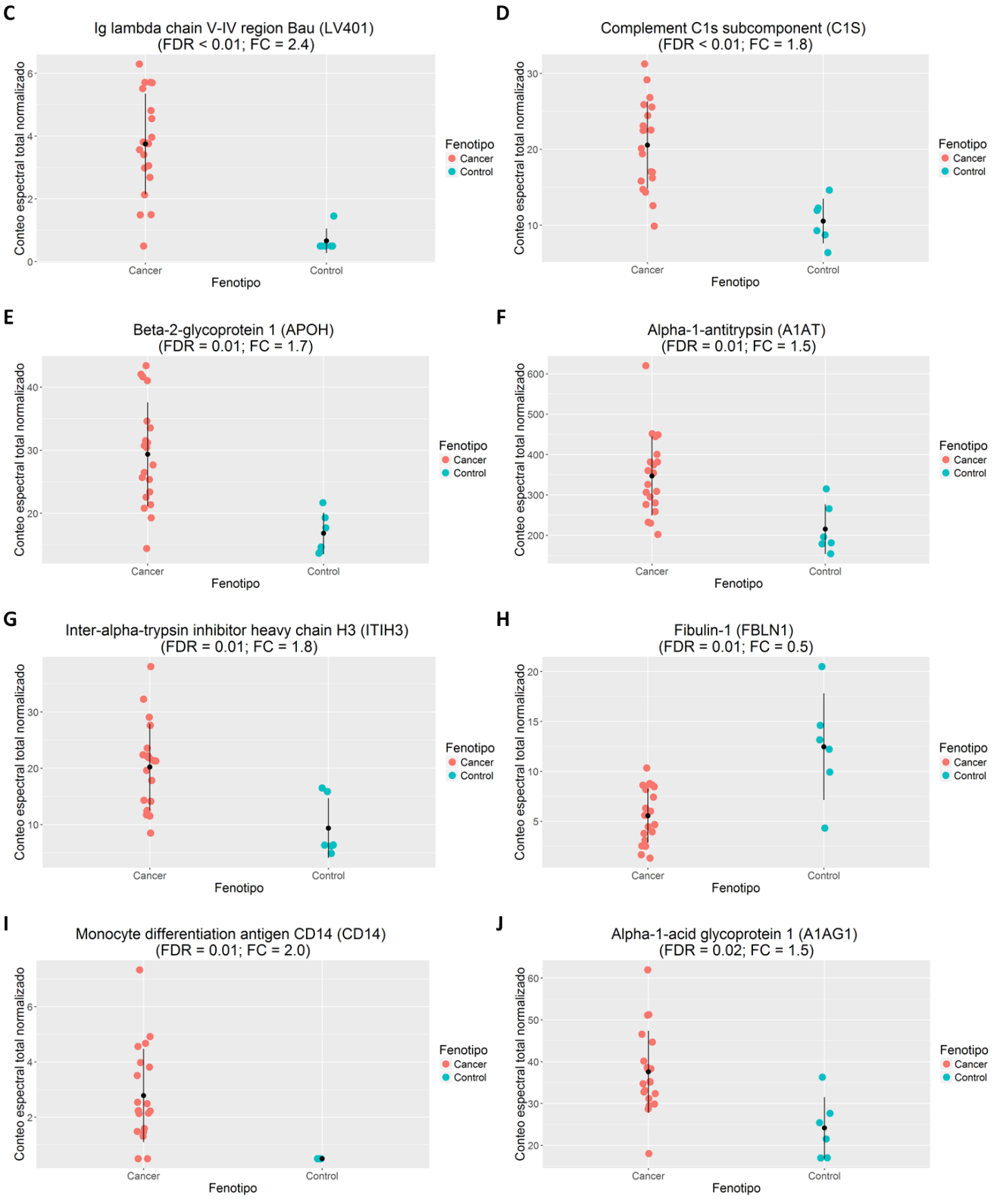
Tabla 27 Proteínas expresadas diferencialmente entre casos de CCR comparado con controles

UNIPROT	ANÁLISIS DE EXPRESIÓN DIFERENCIAL ENTRE CCR Y CONTROLES (Proteínas ~ Fenotipo)				
	Media *	**log2(Fold Change)**	**Fold Change**	***P***	**FDR**
ITIH4_HUMAN	93.64	0.524	1.4	4.5×10^{-6}	7.9×10^{-4}
B4E1F0_HUMAN	44.50	0.553	1.5	2.4×10^{-5}	2.1×10^{-3}
LV401_HUMAN	**2.54**	**1.240**	**2.4**	**3.5×10^{-5}**	**2.1×10^{-3}**
C1S_HUMAN	17.74	0.827	1.8	8.6×10^{-5}	3.8×10^{-3}
APOH_HUMAN	25.96	0.727	1.7	1.4×10^{-4}	0.01
A1AT_HUMAN	315.92	0.630	1.5	1.9×10^{-4}	0.01
ITIH3_HUMAN	17.23	0.866	1.8	4.2×10^{-4}	0.01
FBLN1_HUMAN	**6.65**	**-0.917**	**0.5**	**5.4×10^{-4}**	**0.01**
CD14_HUMAN	**1.76**	**1.011**	**2.0**	**5.0×10^{-4}**	**0.01**
A1AG1_HUMAN	33.98	0.569	1.5	1.5×10^{-3}	0.02
ZA2G_HUMAN	19.44	0.677	1.6	1.5×10^{-3}	0.02
KLKB1_HUMAN	21.60	-0.549	0.7	1.9×10^{-3}	0.03
FIBA_HUMAN	333.15	0.619	1.5	3.5×10^{-3}	0.04
CFAB_HUMAN	76.66	0.343	1.3	3.6×10^{-3}	0.04

Los valores de *P* y FDR corresponden al análisis GLM realizado para evaluar diferencias en la expresión de proteínas entre grupos, y la medida de las diferencias de expresión entre grupos es reportada como log2(*Fold Change*); este valor fue posteriormente usado para calcular el *fold change*. En negrilla se señalan las proteínas con mayores diferencias entre los dos grupos (*fold change* ≥ 2.0 o ≤ 0.5). Las proteínas están listadas en orden ascendente de acuerdo al FDR.
* Media de la expresión de cada proteína en el grupo de controles (referencia)

A continuación se muestran los respectivos gráficos de las diferencias en los niveles de expresión de las 14 proteínas por grupo, inferidos a partir de los conteos espectrales totales normalizados (ver **Figura 14**).




C
Ig lambda chain V-IV region Bau (LV401)
(FDR < 0.01; FC = 2.4)
D
Complement C1s subcomponent (C1S)
(FDR < 0.01; FC = 1.8)
E
Beta-2-glycoprotein 1 (APOH)
(FDR = 0.01; FC = 1.7)
F
Alpha-1-antitrypsin (A1AT)
(FDR = 0.01; FC = 1.5)
G
Inter-alpha-trypsin inhibitor heavy chain H3 (ITIH3)
(FDR = 0.01; FC = 1.8)
H
Fibulin-1 (FBLN1)
(FDR = 0.01; FC = 0.5)
I
Monocyte differentiation antigen CD14 (CD14)
(FDR = 0.01; FC = 2.0)
J
Alpha-1-acid glycoprotein 1 (A1AG1)
(FDR = 0.02; FC = 1.5)
Conteo espectral total normalizado
Fenotipo
Cancer
Control

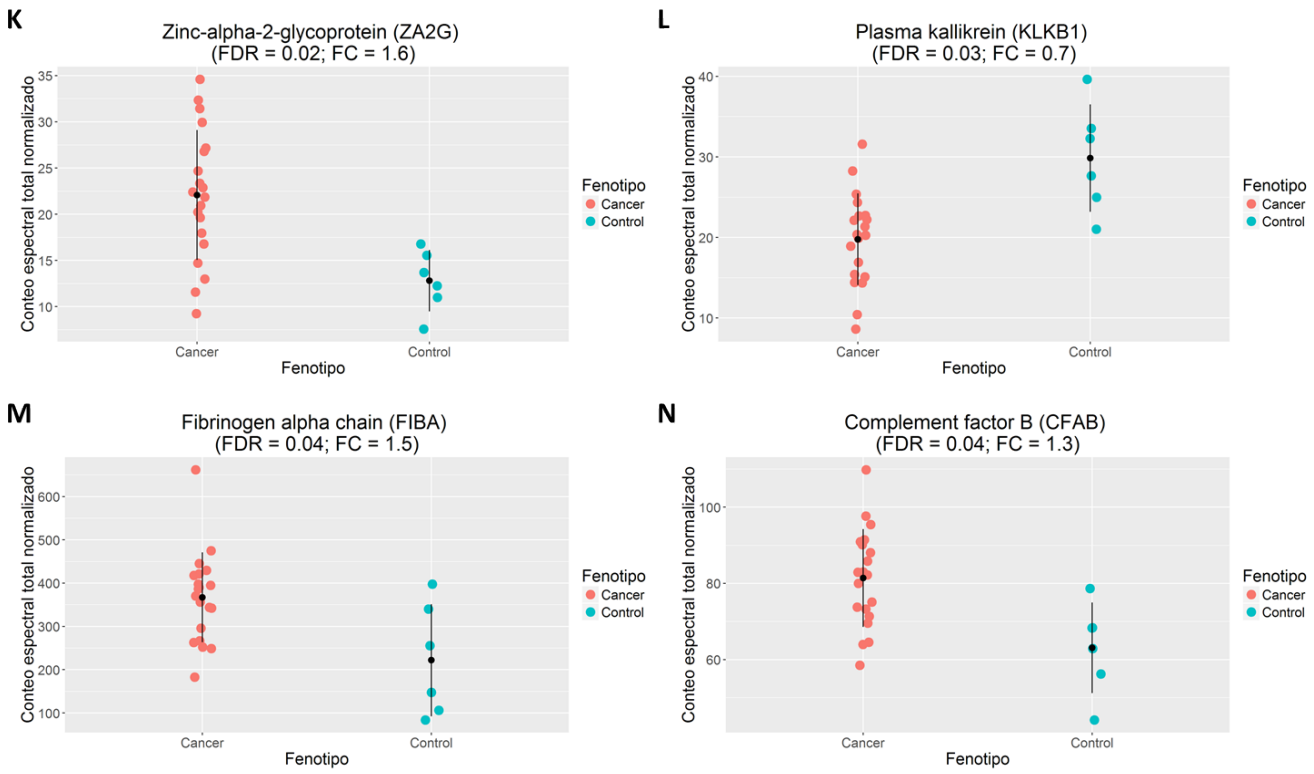


Figura 14 Gráficas de los conteos espectrales normalizados y discriminados por grupos, cáncer o control, de las 14 proteínas identificadas como significativas (FDR < 0.05). Figura original generada en el programa *R statistics* (76).
El grupo cáncer, incluye todos los casos de CCR. El punto negro y las líneas verticales negras, corresponden a la media ± la desviación estándar de la expresión de cada proteína por grupo.
Los nombres de las proteínas y sus respectivas siglas, se mantuvieron en el idioma original de la base de datos Uniprot.
FDR, false discovery rate; FC, *fold change*

Tres de las 14 proteínas identificadas exhibieron diferencias importantes en sus niveles de expresión entre grupos; dentro de éstas las proteínas LV401 y CD14, las cuales se encontraron dos veces más expresadas en casos que en los controles (*fold change* ≥ 2.0), y la FBLN1, que se encontró dos veces menos expresada en los casos que en los controles (*fold change* ≤ 0.5) (ver **Tabla 27** y **Figura 14**).

Selección de proteínas candidatas como posibles biomarcadores en plasma para el control del CCR en Colombia

Teniendo en cuenta los resultados anteriormente expuestos, se evaluó la capacidad discriminatoria de las 14 proteínas diferencialmente expresadas entre los casos de CCR

y controles. Como se puede observar en la **Figura 15**, estas 14 proteínas lograron clasificar adecuadamente las muestras en un análisis no supervisado; solo una muestra control (con código 8023) fue clasificada erróneamente. Por otro lado, usando como clasificadores solo las tres proteínas que además mostraron diferencias importantes en los niveles de expresión entre los dos grupos, LV401, CD14 y FBLN1, se obtuvieron resultados similares al lograr clasificar adecuadamente las muestras en un análisis no supervisado; nuevamente, una muestra control fue clasificada erróneamente (código 68287) (ver **Figura 16**).

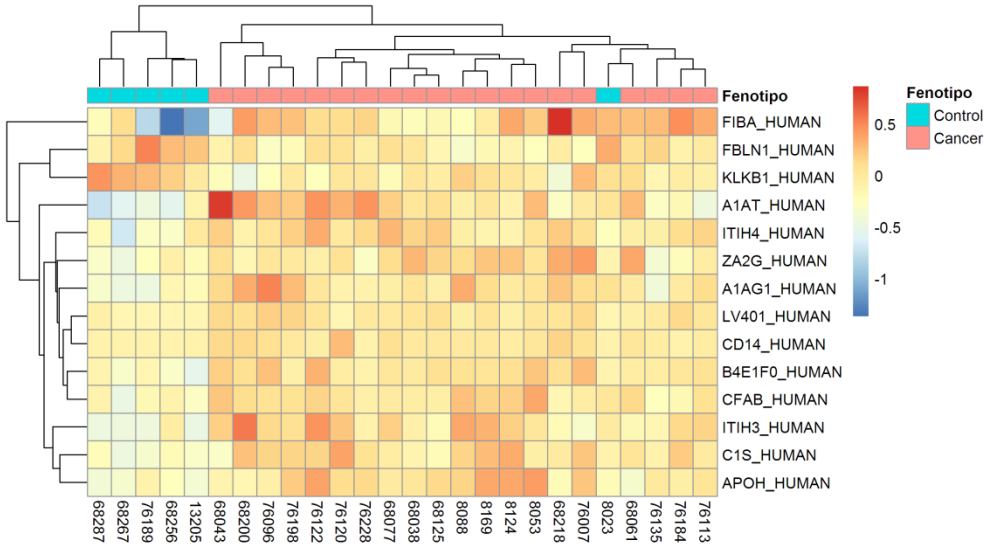


Figura 15 Análisis de clasificación jerárquica no supervisado tipo HeatMap, usando las 14 proteínas con FDR < 0.05. El dendograma de las 14 proteínas con diferencias significativas se muestra a la izquierda; por otro lado, arriba se muestra el dendograma del agrupamiento de las 26 muestras del estudio en dos grandes clusters codificados por colores. Figura original generada en el programa *R statistics* (76).

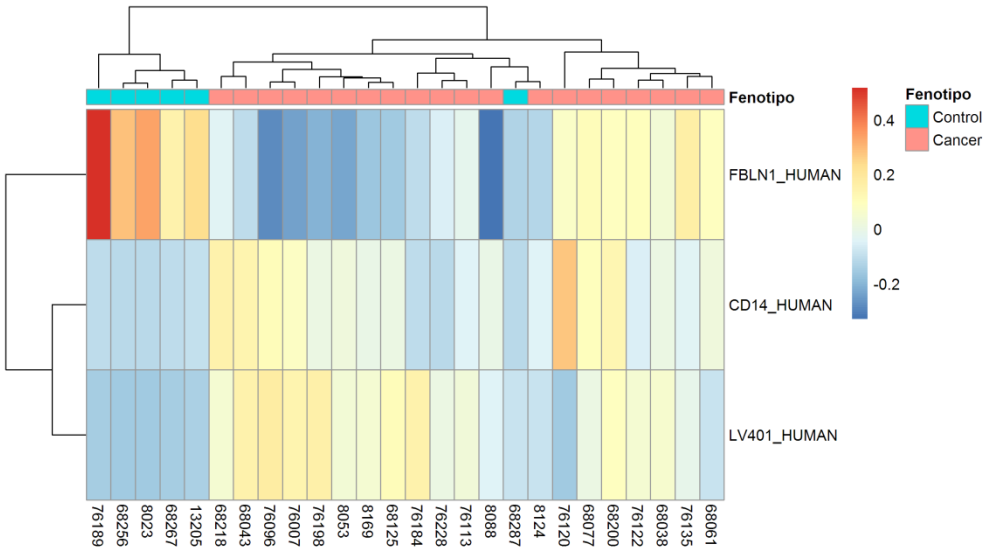
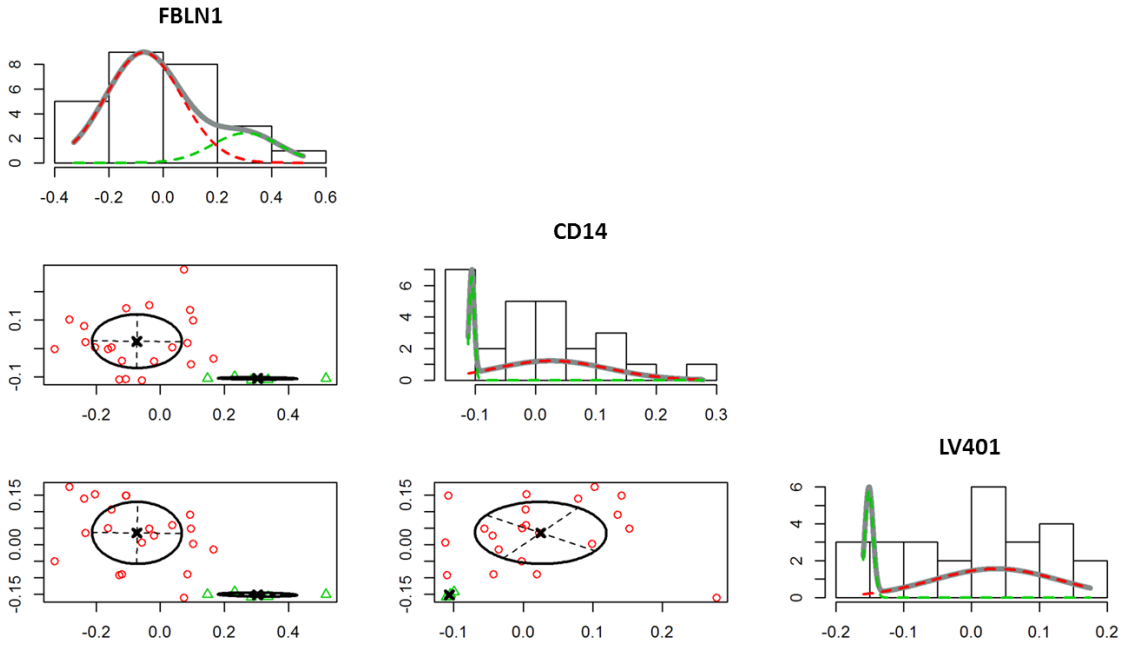


Figura 16 Análisis de clasificación jerárquica no supervisado tipo HeatMap, usando las tres proteínas con FDR < 0.05 y *fold change* ≥ 2.0 o ≤ 0.5. El dendograma de las tres proteínas candidatas se muestra a la izquierda y arriba se muestra el dendograma del agrupamiento de las 26 muestras del estudio en dos grandes clusters codificados por colores. Figura original generada en el programa *R statistics* (76).

En los análisis no supervisados usando un modelo probabilístico con las tres proteínas candidatas, se puede observar una clara separación de las muestras en dos grupos, según los niveles de expresión de FBLN1, CD14 y LV401 (ver **Figura 17**). Los histogramas muestran cómo se distribuye la expresión de cada proteína en los dos grupos formados por el modelo; el eje X corresponde a los conteos espectrales transformados y el eje Y es la proporción de las muestras. Cada histograma, se acompaña de un gráfico tipo PCA que muestra la separación de los grupos usando estos marcadores candidatos. El grupo o cluster 1 está formado por el 81% de la muestra y se caracteriza por una baja expresión de FBLN1 y mayor expresión de CD14 y LV401, con respecto al grupo o clúster 2 (que consiste en el 19% de la muestra). El color rojo corresponde a las muestras agrupadas en el cluster 1, mientras que el color verde hace referencia a las muestras agrupadas en el clúster 2.




```
*********************************************************************
* Number of samples     =  26
* Problem dimension     =  3
*********************************************************************
*        Number of cluster =  2
*               Model Type =  Gaussian_pk_Lk_D_Ak_D
*                Criterion =  BIC(-180.4777)
*               Parameters =  list by cluster
*                   Cluster  1 :
                          Proportion =  0.8077
                               Means =  -0.0723 0.0252 0.0360
                           Variances = |     0.0196    -0.0001    -0.0001 |
                                       |    -0.0001     0.0089    -0.0002 |
                                       |    -0.0001    -0.0002     0.0088 |
*                  Cluster  2 :
                          Proportion =  0.1923
                               Means =  0.3036 -0.1059 -0.1511
                           Variances = |     0.0153    -0.0001    -0.0002 |
                                       |    -0.0001     0.0000     0.0000 |
                                       |    -0.0002     0.0000     0.0000 |
*           Log-likelihood =  116.3036
*********************************************************************
```

Figura 17 Resultados de los análisis no supervisados usando un modelo probabilístico con las tres proteínas candidatas. La parte de arriba de la figura muestra la distribución de los niveles de expresión de cada proteína en histogramas, en donde el eje X corresponde a los conteos espectrales transformados y normalizados, y el eje Y es la frecuencia. Adicionalmente, se muestra la separación de los grupos en gráficos tipo PCAs. En la parte de abajo, se muestra un resumen de los resultados del análisis probabilístico, en su idioma original obtenido en *R statistics* (76). Figura original generada en el programa *R statistics* (76).

Como se observa en la **Tabla 28**, el porcentaje de aciertos en la predicción del fenotipo para las muestras analizadas fue del 96%; éste fue calculado comparando el grupo o cluster que asigna el modelo a cada muestra, con los fenotipos reales. Una muestra control fue clasificada como cáncer (código 68287).

Tabla 28 Porcentaje de aciertos en la predicción del fenotipo usando las tres proteínas candidatas, dentro de un modelo probabilístico no supervisado

N	FENOTIPO		CÓDIGO	PREDICCIÓN
	MODELO PROBABILÍSTICO *	REAL		
1	**1**	**Control**	**68287**	**Errónea**
2	2	Control	76189	Correcta
3	2	Control	68256	Correcta
4	2	Control	68267	Correcta
5	2	Control	13205	Correcta
6	2	Control	8023	Correcta
7	1	Cancer	76120	Correcta
8	1	Cancer	8124	Correcta
9	1	Cancer	8053	Correcta
10	1	Cancer	68218	Correcta
11	1	Cancer	8088	Correcta
12	1	Cancer	68077	Correcta
13	1	Cancer	76096	Correcta
14	1	Cancer	68200	Correcta
15	1	Cancer	76184	Correcta
16	1	Cancer	76122	Correcta
17	1	Cancer	76228	Correcta
18	1	Cancer	68038	Correcta
19	1	Cancer	8169	Correcta
20	1	Cancer	76007	Correcta
21	1	Cancer	68125	Correcta
22	1	Cancer	76135	Correcta
23	1	Cancer	68043	Correcta
24	1	Cancer	76113	Correcta
25	1	Cancer	68061	Correcta
26	1	Cancer	76198	Correcta
ACIERTOS (%)				**96.15**

* Los números 1 y 2 corresponden a los clusters o grupos formados usando las diferencias en la expresión de las proteínas incluidas en el modelo, como se definen en la **Figura 17**.

3.2 Discusión

Actualmente, no existen métodos o biomarcadores de fácil acceso con una alta sensibilidad y especificidad para el tamizaje, diagnóstico o pronóstico del CCR. Si bien la colonoscopia con biopsia es el *gold standard*, en nuestro sistema de salud no está incluida como estrategia de tamizaje poblacional; lo anterior, en parte porque implica un mayor costo y también por su carácter invasivo. Los derivados sanguíneos, plasma y suero, son el tipo de muestras que más predominan en la práctica clínica para el control de enfermedades y, adicionalmente, son el tipo de muestras que más abundan en biobancos, precisamente por su carácter menos invasivo; por lo anterior, resultan ideales para su análisis en estudios de biomarcadores, cuyos candidatos puedan ser posteriormente validados para su uso en la práctica médica (197).

En esta parte del estudio, que consistió en el análisis exploratorio de perfiles proteómicos en plasma de pacientes colombianos con CCR, se usaron técnicas de última generación en proteómica, como es el HPLC acoplado al sistema de MS ESI-Q-Exactive Orbitap. Para esto, se seleccionaron muestras de plasma de 20 casos de CCR y 6 controles que hacían parte del estudio previo de marcadores de susceptibilidad genética del CCR.

En total obtuvimos 469 proteínas en plasma identificadas con mínimo 2 péptidos. Este número se puede considerar bajo, si tenemos en cuenta que en la base de datos del proteoma plasmático (Plasma Proteome Database, PPD; http://www.plasmaproteomedatabase.org/) reportan que se han publicado alrededor de 10546 proteínas plasmáticas en 509 artículos (201). Cabe resaltar que esta diferencia no es del todo inesperada, teniendo en cuenta que en esta misma base de datos reportan grandes variaciones al respecto entre los estudios; pues el 64.12% y el 20.85% se han descrito solo en uno y dos artículos, respectivamente, mientras que un muy bajo porcentaje de éstas (1.70%) están soportadas en más de 10 artículos (201). Esta falta de reproducibilidad entre los estudios, aun con métodos de alta resolución, está en relación con los retos propios de los análisis proteómicos en plasma debido al amplio rango dinámico en la concentración de las proteínas (188). Una alternativa muy costosa, pero muy prometedora por aumentar la resolución y la capacidad de detección de una mayor cantidad de proteínas, es la implementación de métodos en HPLC multidimensionales

acoplados a MS-MS, los cuales se caracterizan por combinar al menos dos métodos de separación cromatográfica; dentro de las combinaciones más usadas se encuentran el acople en línea de un primer sistema de separación por intercambio iónico para fraccionar la muestra compleja según la carga de los analitos, seguido de un segundo sistema de separación de esas fracciones basándose en la hidrofobicidad de los compuestos con un sistema RPLC (175, 176, 202).

A pesar de las limitantes mencionadas anteriormente, se lograron encontrar resultados importantes con respecto a perfiles proteómicos diferenciales entre CCR y controles; esto se evidenció primero en los resultados del PCA que mostraron una adecuada separación de los casos y controles del estudio, usando los dos primeros PCs de la variabilidad proteómica en plasma. Llama la atención que dos muestras controles se mostraron más similares a los casos de CCR; cabe la posibilidad de que estos individuos inicialmente reclutados como *"controles"* correspondan realmente a casos aún no diagnosticados al momento del reclutamiento. Esto no sería de extrañar, debido a que el proceso de carcinogénesis colorrectal toma al menos 10 años (13), periodo en el cual la enfermedad puede ser asintomática; más aún, debido a que la colonoscopia no es un método de tamizaje en Colombia, los controles del estudio se reclutaron cuando reportaron ausencia de antecedentes personales de cáncer y ausencia de síntomas gastrointestinales en la consulta médica de control.

Otro aspecto a mencionar, es la gran variabilidad biológica observada en los casos de CCR, pues están ampliamente distribuidos a lo largo del PC1; éste resulta ser uno de los retos más descritos en los estudios proteómicos, el cual tiene su explicación en que el proteoma de un individuo es dinámico y en que diferentes factores como la edad, el sexo, el estado de ayuno, los factores genéticos y otros, como la comorbilidad, pueden ser fuente de variación en los perfiles de proteínas encontradas en plasma, al ser éste un reflejo del estado fisiopatológico de un individuo (197, 201). Sin embargo, también es importante resaltar que en el proceso de búsqueda de biomarcadores, se espera que estos sean robustos, para que puedan ser implementados a nivel poblacional en el control de las enfermedades de interés, a pesar de la variabilidad biológica entre individuos.

Dentro de los análisis de expresión diferencial de proteínas en solo casos de CCR, no encontramos resultados significativos; lo anterior, pudo deberse al relativamente bajo número de proteínas identificadas en estos análisis, las cuales pueden no ser suficientes para discriminar entre las variedades fenotípicas de una misma enfermedad. Otra razón, es el pequeño tamaño de la muestra usada, lo que se debió en gran medida por la aplicación de diferentes filtros en el proceso de selección de éstas, como fueron la edad, el sexo y el estado de ayuno, y a la limitante en el uso de muestras no autorizadas para estudios futuros.

Por otro lado, logramos encontrar un perfil de 14 proteínas diferencialmente expresadas entre casos de CCR y controles colombianos, todas con un FDR < 0.05. Es de resaltar que varias de estas proteínas encontradas tienen un papel regulador en la matriz extracelular, ECM (ITIH3, ITIH4, KLKB1, A1AT y B4E1F0) o hacen parte de su estructura (FIBA y FBLN1) (203). La ECM consiste en una estructura estable, formada por diferentes compuestos, como proteínas, glicoproteínas, proteoglicanos y polisacáridos, y tiene como funciones principales, servir de soporte estructural y bioquímico entre las células que forman un tejido (204). La ECM es entonces un factor esencial en la regulación de diversos procesos celulares y biológicos, como crecimiento, diferenciación, migración y apoptosis, todos muy importantes en el desarrollo embrionario y, más tarde, en el mantenimiento de la homeostasis tisular para prevenir la transformación neoplásica de los tejidos y apoyar procesos de reparación tisular; lo anterior, se debe a su capacidad de interactuar de forma directa o indirecta con factores de crecimiento como BMPs, FGFs, hedghogs y WNTs, entre otros (203-205). Actualmente, está muy bien documentado el papel de una dinámica anormal en la ECM en cáncer y se reconoce el papel de este componente en el desarrollo de un microambiente tumoral; esto se debe a la alteración en la composición de la ECM la cual puede modificar sus propiedades bioquímicas y arquitectura, favoreciendo los procesos de carcinogénesis que son liderados por las células malignas e inflamatorias (204).

Dos de las proteínas reguladoras de la ECM que encontramos con diferencias significativas, hacen parte de la familia de inhibidores de proteasas, ITI (ITIH3 e ITIH4; Inter-alpha-trypsin inhibitor heavy chain H3 - H4), las cuales se unen covalentemente al ácido hialurónico, HA, y estabilizan la matriz (206, 207). En nuestro estudio, encontramos una sobreexpresión de ambas proteínas en los casos de CCR comparado con los

controles. Otros estudios de casos y controles también han reportado una sobreexpresión en suero o plasma de ambas proteínas en cáncer gástrico (208, 209) y de ITIH4 en pacientes con cáncer de mama (210). Igualmente, en un estudio realizado en modelos animales con deleción de una copia del gen APC (ratas $Apc^{pirc/+}$; susceptibles al desarrollo de tumores colorrectales) versus los tipo silvestre (ratas $Apc^{+/+}$), se encontraron niveles aumentados de ITIH4 en las ratas $Apc^{pirc/+}$, los cuales se correlacionaron tanto con el tiempo como con el número de tumores desarrollados, por lo que los autores la proponen como un biomarcador de diagnóstico temprano del CCR (211). De manera interesante, se ha observado un aumento en los depósitos de HA en los tejidos tumorales sólidos, incluyendo en CCR, lo que se ha correlacionado con un mal pronóstico ya que esto puede ser usado por el tumor como barrera para evadir al sistema inmune; lo anterior, puede explicar la sobreexpresión de ITIH3-4 que nosotros, y otros investigadores, encontramos en pacientes con cáncer (211-213). Sin embargo, otros estudios han propuesto algunas de estas proteínas ITI como supresoras de tumor (207).

La degradación proteolítica de la ECM es necesaria para los procesos de invasión y metástasis tumoral, y se da gracias a la acción de activadores, inhibidores y receptores celulares de diversas proteasas (214, 215). La plasmina, una serina proteasa, y miembros de la familia de las metaloproteasas de matriz, MMPs, juegan un papel clave en la degradación de proteínas estructurales de la ECM (214, 215). La disponibilidad de la plasmina depende de la actividad enzimática de los activadores de plasminógeno urokinasa, uPA, y tisular, tPA, sobre el zimógeno de la plasmina que es el plasminógeno; sin embargo, existe evidencia sobre la actividad proteasa de la kalicreína plasmática, KLKB1 (plasma kallikrein) en la activación del plasminógeno y producción de plasmina (215). Adicionalmente, la KLKB1 también tiene actividad proteolítica sobre kininógenos de alto peso molecular, HMWK, liberando kininas, las cuales tienen funciones en la proliferación celular, la activación de leucocitos y la migración celular, mediante su interacción con el receptor constitutivo 2, B2R, y el receptor inducible 1, B1R; esta vía tiene un gran potencial de estudio en cáncer, puesto que se ha observado que las células tumorales expresan altos niveles de kininas y sus receptores, favoreciendo un mecanismo autocrino para el crecimiento y la progresión tumoral (216). Nosotros encontramos una disminución en la expresión de KLKB1 en el grupo de CCR comparado con el grupo control, lo que está en acuerdo con lo reportado en un estudio reciente por *Peltier J, et al,* quienes observaron niveles bajos de ésta proteína tanto en adenomas

como con CCR, al comprarlo con controles (212). Más aun, otro estudio realizado por *Matsumura Y, et al.,* igualmente observaron una disminución de KLKB1 y HMWK en el plasma de pacientes con diferentes tipos de cáncer comparado con controles, por lo que los autores concluyen que en cáncer aumenta el consumo de estos compuestos para producir kininas (217).

En línea con los anteriores hallazgos, encontramos otra proteína inhibidora de proteasas sobreexpresada en los casos de CCR; esta es la A1AT (alpha-1 antitrypsin) que es una glicoproteína circulante con función inhibitoria sobre la actividad de serinas-proteasas en sangre y tejidos (218). Diferentes estudios han encontrado un aumento en la expresión de ésta proteína en suero de pacientes con CCR y adenomas (212, 218, 219), con cáncer de pulmón y próstata (220) y con leucemia linfoblástica B (221), comparado con controles sanos. Cabe resaltar que, actualmente, se considera que la proteína A1AT puede tener múltiples efectos en el control del crecimiento e invasión tumoral, lo que se ha asociado a la presencia de formas clivadas de la proteína que al parecer muestran diferencias en los niveles de actividad y mecanismos de acción (222, 223).

Con respecto a las glicoproteínas estructurales de la ECM diferencialmente expresadas en nuestro estudio, observamos que FIBA (Fibrinogen alpha chain) o fibrinógeno, estaba aumentada en los pacientes con CCR comparado con los controles. Ésta se encuentra soluble en el plasma sanguíneo y es precursora de la fibrina; ambas importantes en la coagulación y en el monitoreo de la migración celular para la re-epitelización y reparación de heridas (224). Tanto el fibrinógeno como la fibrina se unen al factor de crecimiento de fibroblastos, FGF-2, y modulan sus funciones; esto tiene implicaciones importantes en cáncer, puesto que se ha encontrado que las células tumorales expresan FGF-2 y fibrinógeno de forma endógena, favoreciendo la progresión del tumor (224). Más aun, niveles incrementados de fibrinógeno circulante, se han asociado a un aumento en la formación de complejos entre células tumorales y plaquetas, gracias a que sobreexpresan integrinas tipo $\alpha_v\beta_3$ y $\alpha_{IIb}\beta_3$, respectivamente, que usan como ligando moléculas de fibrinógeno; lo anterior, es un excelente mecanismo de metástasis a distancia de las células tumorales pues favorece la evasión de los sistemas de defensa (213). Su papel en la progresión del tumor está respaldado por estudios recientes que reportan niveles aumentados de fibrinógeno en el suero de pacientes con CCR comparado con pacientes con adenomas (212).

Por el contrario, observamos que la expresión de la proteína estructural de la ECM, FBNL1 (Fibulin1) o fibulina 1, estaba marcadamente disminuida en pacientes con CCR en relación a los controles. Su función principal es estabilizar la ECM a través de su interacción con el receptor de fibronectina y, es considerado como un supresor tumoral debido a su papel en el control de eventos celulares como la adhesión, migración, transformación e invasión (225). Otro estudio realizado en tejido tumoral colorrectal, comparado con su par no tumoral, encontró que la baja expresión de fibulina 1 en los tejidos tumorales se correlacionó con la hipermetilación del promotor del gen *FBLN1* y que estos cambios se asociaron a una menor sobrevida general por la enfermedad (225). Más aun, niveles bajos de fibulina 1 también se han asociado a factores de mal pronóstico de CCR, como el número de ganglios positivos, la presencia de metástasis y estadios avanzados, por lo que la proponen como un marcador de pronóstico (226). Lo anterior se correlaciona con nuestros resultados, teniendo en cuenta que el 75% de los casos de CCR de nuestro estudio son invasivos.

El sistema de complemento es reconocido como un mecanismo innato de defensa que actúa como efector del sistema inmune (227). Si bien se conoce que una de sus funciones es destruir las células tumorales, es posible que algunas de éstas sobrevivan y se transformen, ganando características que les permiten evadir el sistema inmune (227). La B4E1F0 (Plasma protease C1 inhibitor) es una proteína inhibidora de proteasas que tiene como función principal inhibir la activación del sistema de complemento, regulando de ésta manera procesos inflamatorios en el microambiente tumoral (228, 229). Se ha sugerido que las células tumorales pueden sintetizar esta proteína para defenderse del daño celular generado por la activación del sistema de complemento del individuo; lo anterior, con base a que se ha encontrado sobreexpresada en el plasma o suero de pacientes con diferentes tipos de cáncer (229, 230). Igualmente, nosotros encontramos un aumento en la expresión de ésta proteína en los casos de CCR comparado con los controles.

Este mecanismo de defensa de las células tumorales apoyaría la noción de que el sistema de complemento solo tiene un efecto antitumoral; sin embargo, también existe evidencia acerca del papel de éstas proteínas como activadoras o promotoras del crecimiento celular, invasión, angiogénesis, migración y supresión de procesos

antitumorales, favorecidos por el microambiente tumoral (228, 231). En acorde con lo anterior, en nuestro estudio encontramos diferencias significativas en la expresión de proteínas que hacen parte del sistema de complemento (C1S y CFAB) o lo regulan (LV401). La proteína C1S (Complement C1s subcomponent), una serina proteasa que forma parte del complejo C1 y participa en la activación de la vía clásica del sistema de complemento, se ha reportado sobreexpresada a nivel de mRNA y proteína en el tejido de pacientes con cáncer del tracto urinario avanzado (232) y a nivel de proteína en el plasma de hombres con alto riesgo de tener cáncer de próstata (PSA total = 24.6-724 ng/mL) en comparación con otros de bajo riesgo (233), por lo que se ha propuesto como un marcador de pronóstico. Por otro lado, la proteína CFAB (Complement factor B), que hace parte de la vía alterna de activación del sistema de complemento, se ha propuesto como candidato en el diagnóstico de cáncer de páncreas (234). Nuestros resultados están en acuerdo con el papel que se le ha atribuido a éstas dos proteínas en procesos de invasión y migración tumoral (231), pues encontramos niveles elevados de ambas en el grupo de CCR comparado con controles.

Igualmente, nuestros hallazgos con respecto a la asociación entre CCR y niveles aumentados de la proteína regulatoria del sistema de complemento, LV401 (Ig lambda chain V-IV region Bau) o región variable de la cadena ligera lambda de inmunoglobulinas, Ig, están en línea con lo reportado. Si bien se conoce que las Ig son producidas por los linfocitos B activados, y que su región variable de la cadena ligera lambda es esencial en el reconocimiento antigénico requerido para activar la respuesta inmunológica en un individuo, existe nueva evidencia sobre la capacidad de las células tumorales epiteliales para producir y secretar Ig (235, 236). Esta producción endógena de Ig observada en diversas líneas celulares de cáncer y tejidos tumorales, incluyendo CCR, se ha correlacionado con el incremento en la expresión de proteínas asociadas a vías de proliferación y supervivencia celular como Ciclina D1, NF-κβ y PCNA, así como con la disminución en la expresión de proteínas antiapoptóticas como Bcl-2 (235-237).

Todos nuestros hallazgos hasta ahora descritos, apoyan la evidencia científica sobre el papel de diferentes elementos del microambiente tumoral en el acondicionamiento de un entorno favorable para la progresión del CCR. Como se puede deducir, los componentes del sistema inmune son de los más relevantes en el desarrollo de un fenotipo inflamatorio persistente y la malignización del tejido (228). Otro factor muy importante que puede

asociarse con el desarrollo y progresión del CCR, es el establecimiento de un sistema de reconocimiento aberrante de productos de la flora intestinal por parte del receptor de lipopolisacáridos, TLR, de localización transmembranal en las células de defensa y del epitelio intestinal (238). Debido a que en el tracto intestinal existe un flora bacteriana importante, este mecanismo de reconocimiento de endotoxinas y otros productos bacterianos están muy regulados, lo que se ha evidenciado por la baja expresión del complejo CD14/TLR4/MD2 en células epiteliales normales en colon; sin embargo, en CCR y en líneas celulares de CCR se ha evidenciado un aumento en la expresión de TLR4 y en la formación de este complejo (239). Este proceso se ha asociado a la activación de vías proinflamatorias y de proliferación celular como NF-κβ y MAPK, entre otras (238, 239).

En nuestro estudio, encontramos un aumento en la expresión de CD14 mayor al doble en los casos de CCR comparado con los controles. La proteína CD14 (Monocyte differentiation antigen CD14), es un co-receptor de TLR4 (Toll-like receptor 4), que activa el sistema inmune innato (239). Un aumento en la expresión de éste co-receptor puede favorecer un incremento persistente de la respuesta inflamatoria a la flora intestinal y el desarrollo o progresión tumoral. Adicionalmente, se ha observado un número elevado de monocitos circulantes y macrófagos infiltrantes de tumor, TIMs, CD14+ CD169+, en pacientes con CCR, que se han correlacionado con estadios avanzados de la enfermedad (240). Todo lo anterior, apoya nuestros resultados.

Está claro que un aspecto importante de las células tumorales es su alto requerimiento metabólico necesario para mantener las altas tasas de proliferación que las caracteriza; esto incluye una desregulación de la biosíntesis de lípidos (241). Éste incremento en la tasa metabólica lipídica contribuye a diferentes procesos implicados en el desarrollo y progresión del cáncer, como son la producción aumentada de colesterol y lípidos de membrana que contribuyen al acondicionamiento de balsas lipídicas importantes en el tráfico de membrana y la señalización intracelular, así como a la producción de moléculas de señalización como segundos mensajeros, entre otros procesos (241).

Dos de las proteínas diferencialmente expresadas en nuestro estudio, participan en procesos relacionados con la lipólisis y el transporte de lípidos. La proteína ZA2G (Zinc-alpha-2-glycoprotein), que consiste en una proteína soluble que estimula la lipólisis, un

proceso asociado caquexia en pacientes con cáncer, y que puede tener un rol en la respuesta inmune (242); y la proteína APOH (Beta-2-glycoprotein 1), que hace parte de la familia de apolipoproteínas, las cuales tienen como función formar lipoproteínas para el transporte de lípidos (243). En los dos casos, encontramos niveles aumentados de éstas en el grupo de CCR, comparado con los controles, reflejando precisamente un aumento en la tasa metabólica lipídica en cáncer; nuestros resultados están en acuerdo con otros estudios en diferentes tipos de tumores malignos que los han propuesto como marcadores de pronóstico (243, 244).

Podemos decir que hemos encontrado proteínas diferenciales, en su mayoría, con potencial como biomarcadores de pronóstico según la literatura. Sin embargo, identificamos una proteína con potencial como biomarcador predictivo; ésta es la proteína A1AG1 (Alpha-1-acid glycoprotein 1), que funciona como una proteína transportadora y moduladora del sistema inmune, pero que además se une a drogas sintéticas, influenciando su distribución y disponibilidad en el organismo (245). Niveles aumentados de ésta proteína se han encontrado en pacientes con CCR progresivo no respondedores al tratamiento con 5-fluoracilo, y también en pacientes con cáncer de pulmón no respondedores al tratamiento con docetaxel; lo anterior, posiblemente por una disminución en la fracción libre de éstos medicamentos, afectando su actividad en el control del tumor (245, 246). En nuestro estudio observamos una mayor expresión de A1AG1 en los casos de CCR que en los controles, sin embargo no podemos evaluar su papel como predictor de la respuesta al tratamiento, pues no tenemos acceso a esta información.

Todas las 14 proteínas mencionadas, permitieron diferenciar los casos de los controles en un análisis jerárquico no supervisado; solo una muestra control fue agrupada con los casos, como se observó en el PCA. De éstas, tres mostraron diferencias importantes en los niveles de expresión entre los dos grupos (*fold change* ± 2 veces), la LV401, la CD14 y la FBLN1; y su capacidad para diferenciar los casos de los controles se pudo evidenciar en los análisis no supervisados con un acierto del 96%. Nuevamente, un muestra control fue agrupada con los casos, lo que fue evidente en el PCA; como se mencionó previamente, es posible que éstas muestras inicialmente incluidas como *“controles”*, correspondieran a individuos con alguna lesión premaligna o maligna aun no diagnosticada y en curso.

En general, nuestros resultados son comparables con los reportados recientemente por *Peltier J, et al*, quienes realizaron un análisis HPLC acoplado a ESI-LQT-Orbitrap, para evaluar diferencias en el proteoma de suero de pacientes con adenoma y adenocarcinoma colorrectales, y controles (212). Ellos identificaron un total de 348 proteínas y seleccionaron 89 como diferencialmente expresadas entre grupos, en base a un valor de *P* no corregido ≤ 0.05 y un *fold change* de ± 1.5 veces; de éstas, en su gran mayoría correspondieron a proteínas regulatorias con actividad enzimática, como proteínas de la familia de SERPINAS, incluyendo la A1T1, la cual proponen como marcador de pronóstico por su papel en procesos de migración y exponen la utilidad de estudiar su expresión en estadios tempranos de la enfermedad (212).

En nuestro estudio que incluyó una técnica de HPLC/MS-MS similar, logramos identificar 469 proteínas con una alta confiabilidad; sin embargo, al aplicar métodos estadísticos más estrictos para la selección de las proteínas diferencialmente expresadas solo nos quedamos con 14 de éstas. Es de resaltar que todas éstas reflejan diferentes procesos de malignización que se han descrito antes para CCR, en una dirección concordante con lo reportado en la literatura.

Dentro de las limitantes están que no logramos ver diferencias entre los casos de CCR in situ comparado con los invasivos, y que no tenemos información sobre el estadio de la enfermedad o la respuesta al tratamiento; por lo anterior, no podemos concluir a partir de nuestros resultados, la utilidad real de estos marcadores para el diagnóstico temprano, pronóstico o su valor predictivo. Sin embargo, de acuerdo a la literatura se podría proponer que ITIH4, KLKB1 y A1T1, tienen potencial como marcadores de diagnóstico temprano al haber sido detectadas diferencialmente en adenomas en humanos o en modelos animales; mientras que la proteína A1AG1 puede tener potencial como marcador predictivo de la respuesta a la quimioterapia. Todas las demás proteínas identificadas, incluyendo LV401, la CD14 y la FBLN1, pueden tener potencial como biomarcadores de pronóstico pues se han asociado a estadios avanzados de cáncer.

Podemos concluir de éste capítulo que se logró identificar un perfil de 14 proteínas diferencialmente expresadas entre casos de CCR comparado con controles colombianos, tres de las cuales son fuertes candidatos como biomarcadores en CCR, posiblemente

para uso en la evaluación del pronóstico de la enfermedad, teniendo en cuenta lo reportado en la literatura. Es de resaltar que se requieren estudios adicionales de validación en un mayor número de muestras de pacientes con CCR en diferentes estadios de la enfermedad, con el fin de evaluar adecuadamente su utilidad clínica.

4. Conclusiones, perspectivas, productos, pasantías y premios

4.1 Conclusiones

i) Existen diferencias en las proporciones de ancestría europea, amerindia y africana en las seis poblaciones colombianas seleccionadas en nuestro estudio, y estas diferencias son el resultado del patrón de mestizaje que se llevó a cabo en estas regiones (sesgos de género), el cual estuvo condicionado por el modelo migratorio de las poblaciones continentales que se asentaron a lo largo del territorio nacional y el rol que éstos ancestros asumieron durante el periodo de la colonización en Colombia.

ii) La ancestría europea se asoció al riesgo de PA en los colombianos, mientras que la ancestría africana se asoció al riesgo de PA y CCR, aún después de ajustar los análisis por otras variables como el nivel educativo (indicador de estrato socioeconómico) y el consumo de AINES, entre otros. Estos resultados son de gran importancia, teniendo en cuenta que Cali en Valle del Cauca no solo es la ciudad con mayor porcentaje de afrocolombianos del país, sino que es de las que se han evidenciado con mayor riesgo de mortalidad por CCR. La identificación de poblaciones de alto riesgo para la enfermedad, tiene implicaciones muy relevantes a nivel de salud pública.

iii) Logramos identificar dos nuevas variantes genéticas de riesgo para el desarrollo de tumores colorrectales en colombianos, no antes reportadas. Por sus características, estos SNPs podrían ser funcionales lo que puede ser investigado en ensayos futuros.

iv) En los análisis proteómicos, identificamos un perfil de 14 proteínas diferencialmente expresadas entre casos de CCR comparado con controles colombianos. Tres de estas proteínas son fuertes candidatos como biomarcadores en CCR, posiblemente para uso en la evaluación del pronóstico de la enfermedad, teniendo en cuenta lo reportado en la literatura.

4.2 Perspectivas

i) Analizar las diferencias de los componentes ancestrales en marcadores uniparentales de nuestros controles, con el fin complementar la información obtenida en esta tesis sobre las estimaciones con los marcadores biparentales.

ii) Implementar métodos de mapeo por mestizaje para identificar loci específicos de ancestría asociadas al riesgo de tumores colorrectales en colombianos. Esto, posiblemente en asocio con otros grupos de investigación en genética del CCR, con los cuales sea posible aumentar el tamaño de la muestra.

iii) Realizar ensayos de los SNPs encontrados como asociados para evaluar si son o no funcionales. Por ejemplo, para el SNP en *TEP1* se puede plantear un ensayo para evaluar la longitud de los telómeros en linfocitos periféricos y comparar entre casos y controles, y según el genotipo. Igualmente, para el SNP en *TK1*, se pueden plantear ensayos de proliferación celular alelo-específicos. Esto lo estamos coordinando actualmente con nuestros colaboradores en la Universidad de Louisiana.

iv) Validar los SNPs que encontramos en nuestro estudio en una muestra independiente colombiana. Al respecto estamos en conversaciones con un grupo par para realizar esta actividad.

v) Validar las proteínas que encontramos diferencialmente expresadas entre grupos, y otras relacionadas a las vías de señalización en las que participan, en una muestra independiente de casos y controles, mediante inmunoensayos de fácil implementación clínica. Al respecto, podemos contar con las muestras biológicas del Biobanco de Tumores del Instituto Nacional de Cancerología; esto, bajo la aprobación de un proyecto de investigación.

4.3 Productos

i) Manuscritos publicados

María Carolina Sanabria-Salas, Gustavo Hernández-Suárez, Adriana Umaña-Pérez, Konrad Rawlik , Albert Tenesa, Martha Lucía Serrano López , Myriam Sánchez De Gómez , Martha Patricia Rojas , Luis Eduardo Bravo , Rosario Albis , José Luis Plata , Heather Green , Theodor Borgovan , Li Li , Sumana Majumdar , Jone Garai , Edward Lee, Hassan Ashktoab , Hassan Brim , Li Li , David Margolin , Laura Fejerman , Jovanny Zabaleta, **"*IL1B-CGTC* haplotype is associated with colorectal cancer in admixed individuals with increased african ancestry".** SCIENTIFIC REPORTS. **2017**, 7:41920, DOI: 10.1038/srep41920

Ruth Andrea Rodríguez, Wilmer Alexis Urrego, María Carolina Sanabria, Myriam Sánchez-Gómez, Adriana Umaña-Pérez, **"Implementación de una metodología para la separación de proteomas de plasma humano mediante electroforesis bidimensional"**. REV. COLOMB. QUIM. **2015**, 44(3), 30-38.

Gustavo Adolfo Hernández Suarez, María Carolina Sanabria Salas, Martha Lucia Serrano López, Oscar Fernando Herrán Falla, Jovanny Zabaleta, Albert Tenesa, **"Genetic ancestry is associated with colorectal adenomas and adenocarcinomas in latino populations".** EUROPEAN JOURNAL OF HUMAN GENETICS. **2014**, V.22 FASC.N/A P.1208 – 1216

María Carolina Sanabria Salas, Yadi Adriana Umaña Pérez, Martha Lucia Serrano, Myriam Sánchez De Gómez, Jorge Meza, Gustavo Hernández Suarez, **"Vías de carcinogénesis colorrectal y sus implicaciones clínicas".** REVISTA COLOMBIANA DE CANCEROLOGÍA. **2012**, V.16 FASC.3 P.170 – 181, 2012

ii) Ponencias orales

NACIONAL: María Carolina Sanabria-Salas, Adriana Umaña-Pérez, Konrad Rawlik, Albert Tenesa, Martha Lucía Serrano-Pérez, Myriam Sánchez De Gómez, Martha Patricia Rojas, Jovanny Zabaleta, Gustavo Adolfo Hernández-Suárez, **"Variantes genéticas comunes en el gen *TEP1* y en el gen *TK1* están asociadas al riesgo de tumores colorrectales en Latinos".** V JORNADAS DE INVESTIGACIÓN EN CÁNCER **2016**; OCT 13-15; CALI, VALLE DEL CAUCA.

INTERNACIONAL: María Carolina Sanabria-Salas, Gustavo Adolfo Hernández-Suárez, Adriana Umaña-Pérez, Martha Lucía Serrano-Pérez, Myriam Sánchez De Gómez, Martha Patricia Rojas, Jovanny Zabaleta, Konrad Rawlik And Albert Tenesa, **"Abstract pr08: The role of genetic structure in colombian coastal and andean populations on disparities in colorectal adenomas and cancer risk".** Proceedings of the EIGHTH AACR CONFERENCE ON THE SCIENCE OF HEALTH DISPARITIES IN RACIAL/ETHNIC MINORITIES AND THE MEDICALLY UNDERSERVED; NOV 13-16, **2015**; ATLANTA, GA. CANCER EPIDEMIOL BIOMARKERS PREV 2016; 25(3 SUPPL):ABSTRACT NR PR08. (Presentación oral y poster)

NACIONAL: María Carolina Sanabria-Salas, ***"Epidemiología Molecular en Cáncer"***. JORNADAS DE INVESTIGACIÓN INSTITUCIONAL, INC, 2015; SEP 09-10; BOGOTÁ, D.C.

NACIONAL: María Carolina Sanabria-Salas, **"Factores de riesgo ambientales y genéticos para cáncer colorrectal en Colombia".** IV JORNADA DE INVESTIGACIÓN EN CÁNCER **2013**; SEP 13-14; MEDELLÍN, ANTIOQUÍA.

NACIONAL: María Carolina Sanabria-Salas, ***"Polimorfismos genéticos asociados al cáncer colorrectal en poblaciones colombianas"***. III JORNADAS DE INVESTIGACIÓN EN CÁNCER 2011; BOGOTÁ D.C.

iii) Posters

María Carolina Sanabria-Salas, Ruth Rodríguez-Castro, Martha Lucía Serrano-López, Gustavo Hernández-Suárez, Myriam Sánchez de Gómez, Adriana Umaña-Pérez. ***"Differential plasma proteome analysis in patients with colorectal cancer from Colombia"*.** ANNUAL MEETING 2017 - AMERICAN ASSOCIATION FOR CANCER RESEARCH; APR 01-05, **2017**; WASHINGTON DC.

María Carolina Sanabria-Salas, Jovanny Zabaleta, Adriana Umaña-Pérez, Konrad Rawlik, Albert Tenesa, Martha Lucía Serrano-López, Myriam Sánchez de Gómez, Martha Patricia Rojas, Luis Eduardo Bravo, Gustavo Hernández-Suárez. ***"Common genetic variants within TEP1 gene (14q11.2 locus) and TK1 gene (17q25.3 locus) are associated with the risk of colorectal tumors in Latinos".*** ANNUAL MEETING 2017 - AMERICAN ASSOCIATION FOR CANCER RESEARCH; APR 01-05, **2017**; WASHINGTON DC.

María Carolina Sanabria-Salas, Gustavo Adolfo Hernández-Suárez, Adriana Umaña-Pérez, Martha Lucía Serrano-Pérez, Myriam Sánchez De Gómez, Martha Patricia Rojas, Jovanny Zabaleta, Konrad Rawlik And Albert Tenesa, **"Abstract pr08: The role of genetic structure in colombian coastal and andean populations on disparities in colorectal adenomas and cancer risk".** Proceedings of the EIGHTH AACR CONFERENCE ON THE SCIENCE OF HEALTH DISPARITIES IN RACIAL/ETHNIC MINORITIES AND THE MEDICALLY UNDERSERVED; NOV 13-16, **2015**; ATLANTA, GA. CANCER EPIDEMIOL BIOMARKERS PREV 2016; 25(3 SUPPL):ABSTRACT NR PR08. (Presentación oral y poster)

María Carolina Sanabria-Salas, Gustavo Adolfo Hernandez Suarez, Yadi Adriana Umana Perez, Martha Lucia Serrano Lopez, Myriam Sanchez De Gomez, Martha Patricia Rojas, Jovanny Zabaleta, ***"Abstract b23: Colombians with high percentage of african ancestry, carrying a specific IL1B haplotype, have increased risk of colorectal***

cancer". SIXTH AACR CONFERENCE ON THE SCIENCE OF CANCER HEALTH DISPARITIES IN RACIAL/ETHNIC MINORITIES; DIC 06-09, **2013**; ATLANTA, GEORGIA. *CANCER EPIDEMIOLOGY BIOMARKERS AND PREVENTION. ISSN: 1055-9965 ED: AMERICAN ASSOCIATION FOR CANCER RESEARCH. V.23 FASC.B23 P.1538 - 7755, 2014*

María Carolina Sanabria-Salas, Yadi Adriana Umaña Pérez, Martha Lucia Serrano, Myriam Sánchez De Gómez, Martha Patricia Rojas, Jovanny Zabaleta, Gustavo Adolfo Hernández Suarez, ***"Association between IL1B gene haplotype and colorectal cancer risk in Colombian populations: a case control study".*** ANNUAL MEETING 2013 - AMERICAN ASSOCIATION FOR CANCER RESEARCH; MAR 06-10, **2013**; WASHINGTON DC.

María Carolina Sanabria-Salas, Yadi Adriana Umaña Pérez, Martha Lucia Serrano, Myriam Sánchez De Gómez, Martha Patricia Rojas, Jovanny Zabaleta, Gustavo Adolfo Hernández Suarez, ***"Abstract b8: Sex-specific association between IL1B-511 polymorphism and colorectal cancer in colombian populations".*** SECOND AACR INTERNATIONAL CONFERENCE ON FRONTIERS BASIC CANCER RESEARCH; SEP 14-18, **2011**; SAN FRANCISCO, CA. *CANCER RESEARCH ISSN: 1538-7445 ED: V.71 FASC.18 SUPPL P.B8 - B8, 2011*

4.4 Pasantías

2015 – 2015; Pasantía en análisis estadístico y bioinformático de datos genómicos usando R statistics, ShapeIT y RFMix (estimación de ancestría local a partir de datos de genoma completo) (4 meses). Genetics and Genomics Department, Roslin Institute, University of Edinburgh, UK. Asesor: Albert Tenesa, PhD.

2014 – 2014; Curso de extensión: Wellcome Trust / EBI Workshop; *"Proteomics Bioinformatics"* (6 días). Noviembre 10-14, 2014, Wellcome Genome Campus, Hinxton, Cambridge, UK.

2013 – 2014; Pasantía en HPLC/MS-MS con la tecnología Orbitrap (3 meses). Proteomics Core, UCDavis Genome Center, Davis, CA, US. Asesor: Servicio de Proteómica de UCDavis; Brett Phinney, PhD.

2012 – 2012; Curso de extensión: Wellcome Trust Advance Course; Human Genome Analysis: Genetic Analysis of Multifactorial Diseases (7 días). Julio 11-17, 2012, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

2011 – 2011; Pasantía en tecnología de microarreglos para genotipar a gran escala (6 meses). Stanley S Scott Science Center, Louisiana State University, New Orleans, LA, US. Asesor: Jovanny Zabaleta, PhD.

4.5 Premios

American Association for Cancer Research: *"Eighth AACR Conference on the Science of Cancer Health Disparities in Racial/Ethnic Minorities and the Medically Underserved"*, Nov 13–16, **2015**, Atlanta, GA, USA. Scholar-in-Training Award ($ 2,750 USD).

EMBL-EBI / Wellcome Trust Course: *"Proteomics Bioinformatics"*, Nov 10-14, **2014**, Wellcome Genome Campus, Hinxton, Cambridge, UK. Beca del 50% para asistir al curso.

American Association for Cancer Research: *"Sixth AACR Conference: The Science of Cancer Health Disparities in Ethnic Minorities and the Medically Underserved"*, Diciembre 6-9, **2013**, Atlanta, GA, USA. Scholar in-Training Award ($ 1,000 USD).

Wellcome Trust Sanger Institute – Advanced Courses: *"Human Genome Analysis: Genetic Analysis of Multifactorial Diseases"*, Julio 11-17, **2012**, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. Beca del 50% para asistir al curso.

A. Anexo 1 - Materiales y métodos en análisis genéticos

Población de estudio

Este es un estudio multicéntrico de casos y controles que incluyó 313 casos de CCR, 200 casos de PA y 506 controles originarios de seis (6) ciudades colombianas de la región andina (Bogotá y Bucaramanga) y costera (Cali en el Pacífico, y Cartagena, Santa Marta y Barranquilla en el Caribe), con diferentes tasas de incidencia y mortalidad por CCR, para un total de 1019 individuos. Todos los casos eran incidentes, confirmados con histopatología, que no habían recibido tratamiento con quimio o radio terapia al momento de la captura (desde el 2008 al 2011). Los controles consistieron en individuos que asistieron a consulta de medicina general por motivos diferentes a síntomas gastrointestinales. Ningún participante estaba relacionado con otro del estudio y ninguno reportó antecedentes personales de otros cánceres. El rango de edad de los participantes fue entre los 30 a 79 años, pues el enfoque del estudio es en CCR de tipo esporádico.

Este trabajo fue aprobado como un *"estudio de riesgo mayor al mínimo"* por el Comité de Ética del Instituto Nacional de Cancerología de Bogotá, Colombia, según las guías para la investigación en humanos dispuestas en la Resolución 8430 de 1993 (De los Aspectos Éticos de la Investigación en Seres Humanos, Título II, Capítulo 1) publicado por el Ministerio de Salud, Republica de Colombia (https://www.invima.gov.co/images/pdf/medicamentos/resoluciones/etica_res_8430_1993.pdf), razón por la cual todos los participantes firmaron un consentimiento informado previo a la recolección de información epidemiológica y de muestras biológicas.

Extracción del DNA genómico

El DNA fue extraído a partir de 200µl de la fracción de células blancas de sangre periférica (buffy coat), usando el kit QIAamp DNA Blood Mini Kit (QIAGEN, Valencia, CA) según recomendaciones del fabricante; se resuspendió en 100µl de agua libre de nucleasas (*Ambion*, Foster *City*, CA) y se almacenó a -20°C. La pureza y concentración del DNA se evaluó en un NanoDrop 2000 (Thermo Scientific, Wilmington, DE). En todos los casos la razón 260/280 fue de ~1.8.

Genotipificación a mediana/gran escala y control de calidad

Se usaron dos plataformas de Illumina® para estudiar variantes genéticas tipo polimorfismos de un solo nucleótido (SNPs): i) un microarreglo de genes candidatos (CG) que incluye 1421 SNPs, conocido como "*Cancer SNP Panel",* y ii) un microarreglo de genoma completo (GWAS) que incluye 958178 SNPs, conocido como "*Infinium*® *OmniExpressExome Array"*. Lo anterior, según las recomendaciones de los fabricantes.

Debido a las diferencias en el número de SNPs incluidos en las plataformas usadas para genotipar las muestras del estudio, los pasos de control de calidad (por SNPs y por muestras) se hicieron por separado, siguiendo las recomendaciones descritas en el protocolo de *Anderson CA, et al (117)*, usando los programas PLINK v1.9 (119) and R statistics v3.2.2 (76).

Control de calidad de la base de datos CG: esta base de datos consistió en genotipos de 1421 SNPs en un total de 521 muestras. Se excluyeron 184 SNPs únicos por presentar diferencias significativas en errores de genotipado entre casos y controles ($P < 0.01$; n = 19), por tener una frecuencia del alelo menor, MAF < 0.04 (n = 42), por presentar una tasa de error en el genotipado > 0.05 (n = 48) o por estar en desequilibrio de Hardy-Weinberg en los controles ($P < 0.01$; n = 98). Se excluyeron 38 muestras únicas por presentar una tasa de error en el genotipado ≥ 0.03 y/o por tener tasas de heterocigocidad > 3 desviaciones estándar de la media (SD) (n = 26) (ver **Figura - Anexos A**) o por tener un valor de identidad por descendencia (IBD) > 0.35 entre par de muestras, en cuyo caso se eliminó una de cada par con el fin de evitar muestras duplicadas, contaminadas o consanguíneas en segundo grado (n = 12). Debido a que

este panel solo incluía 13 SNPs en el cromosoma X, no se aplicó en este caso el filtro por discordancia de sexo/género sino que se tomó el registrado en las bases de datos. La base de datos limpia y final para posteriores análisis quedó con 1237 SNPs (ver **Figura - Anexos B**) en 483 muestras (ver **Tabla - Anexos A**).

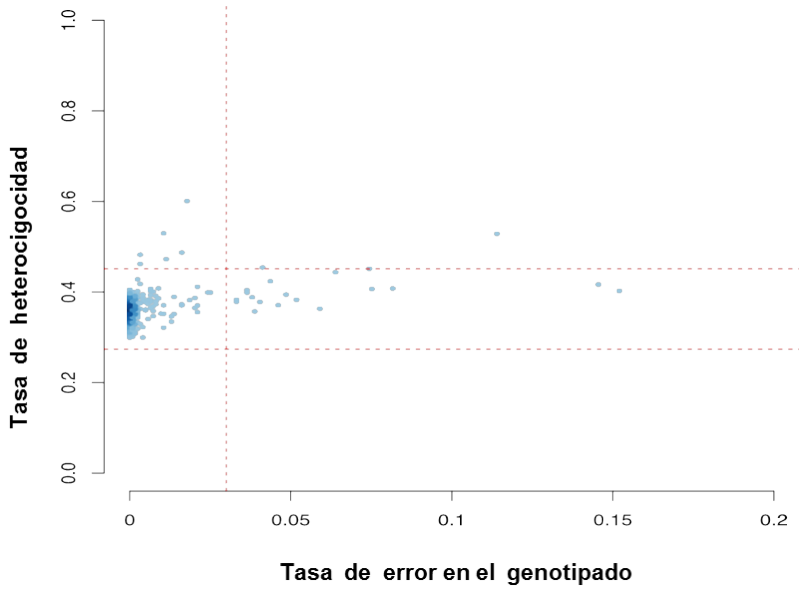


Figura - Anexos A Control de calidad por muestras de la base de datos CG. Se observan las muestras que se eliminaron usando los filtros de tasa de error en el genotipado ≥ 0.03 y/o tasas de heterocigocidad > 3 SD (n = 26). Figura original generada en el programa *R statistics* (76).

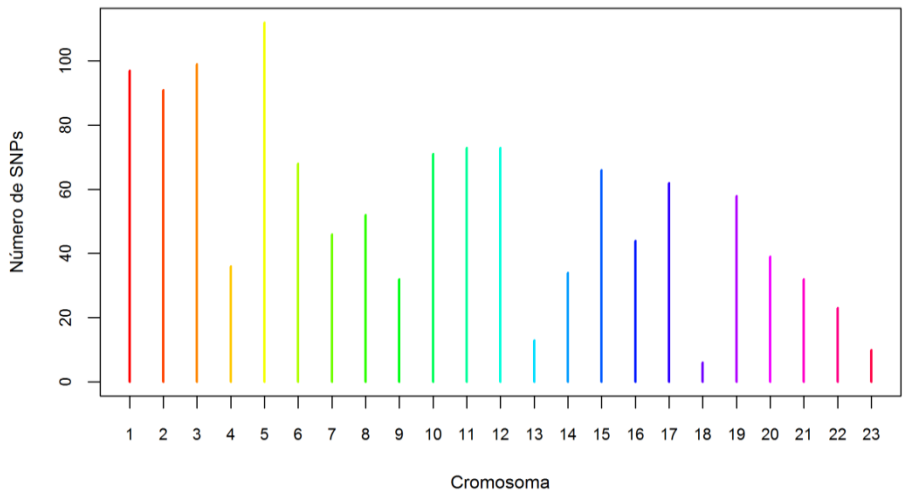


Figura - Anexos B Densidad de SNPs en la base de datos CG limpia. Se observa el número de SNPs por cada cromosoma una vez realizados los pasos de control de calidad por SNPs (n = 1237). Figura original generada en el programa *R statistics* (76).

Tabla - Anexos A Distribución de las muestras que pasaron los controles de calidad

CARACTERÍSTICAS	SETS DE MUESTRAS COLOMBIANAS GENOTIPADAS	
	Plataforma "CG"	Plataforma "GWAS"
Número de muestras*	483	415
Sexo		
Femenino	258	211
Maculino	225	204
Fenotipo		
Control	218	131
PA	101	122
CCR	164	162
Edad		
Mínimo	30.00	30.00
1er cuartil	47.00	52.00
Mediana	56.00	60.00
Media	55.52	58.72
3er cuartil	64.50	66.00
Máximo	74.00	79.00
Ciudad de origen		
Bogotá D.C.	70	115
Bucaramanga	134	114
Cali	58	42
Barranquilla	85	61
Cartagena	71	45
Santa Marta	65	38
Región		
Andina	204	229
Costera	279	186
Nivel educativo		
Sin educación	19	13
Primaria	174	176
Secundaria	171	136
Técnico	45	37
Universidad o mayor	74	52
Sin dato	0	1
Historia familiar de CCR		
No	307	251
Sí	176	164
Consumo de AINES		
No	374	316
Sí	109	99

* Muestras que pasaron los controles de calidad aplicados a cada base de datos, CG y GWAS, por separado. Entre los dos sets de muestras, 85 están repetidas y fueron usadas para hacer las correlaciones en las estimaciones de ancestría obtenidas con diferentes metodologías. En total, son 813 muestras únicas con genotipos a mediana o gran escala, sobre los cuales se realizaron

posteriormente los análisis de regresión logística para evaluar las asociaciones entre la ancestría genética o SNPs con el riesgo de tumores colorrectales.

Control de calidad de la base de datos GWAS: esta base de datos consistió en genotipos de 958178 SNPs en un total de 443 muestras. Se excluyeron 237363 SNPs únicos por presentar diferencias significativas en errores de genotipado entre casos y controles (P < 0.00001; n = 47), por tener un MAF < 0.01 (n = 224698), por presentar una tasa de error en el genotipado > 0.05 (n = 14107) o por estar en desequilibrio de Hardy-Weinberg en los controles (P < 0.00001; n = 187). Se excluyeron 28 muestras únicas por presentar una tasa de error en el genotipado ≥ 0.03 y/o por tener tasas de heterocigocidad > 3 SD (n = 8) (ver **Figura - Anexos C**), por tener un valor IBD > 0.185 entre par de muestras, en cuyo caso se eliminó una de cada par con el fin de evitar muestras duplicadas, contaminadas o consanguíneas en primer y segundo grado (n = 1), o por discordancia entre el sexo/género genético en comparación al registrado (n = 22). La base de datos limpia y final a ser usada en análisis posteriores quedó con 720815 SNPs (ver **Figura - Anexos D**) en 415 muestras (ver **Tabla - Anexos A**).

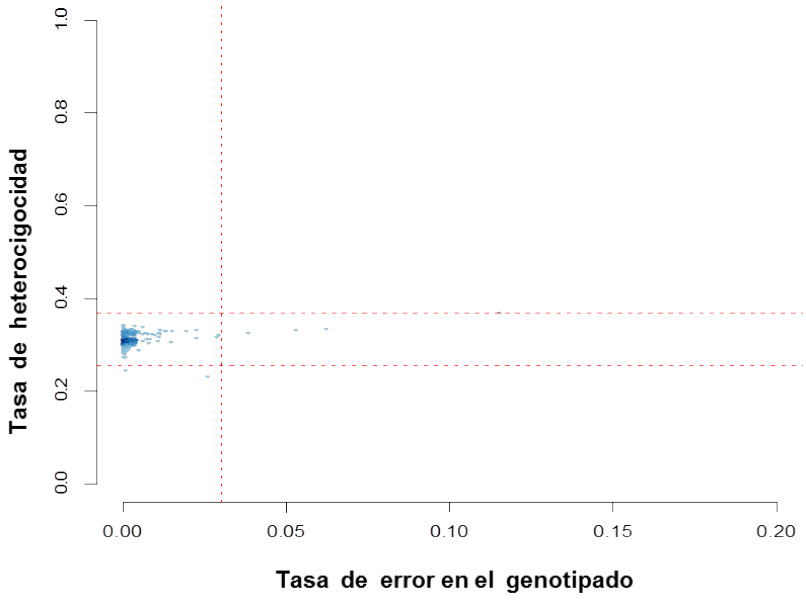


Figura - Anexos C Control de calidad por muestras de la base de datos GWAS. Se observan las muestras que se eliminaron usando los filtros de tasa de error en el genotipado ≥ 0.03 y/o tasas de heterocigocidad > 3 SD (n = 8). Figura original generada en el programa *R statistics* (76).

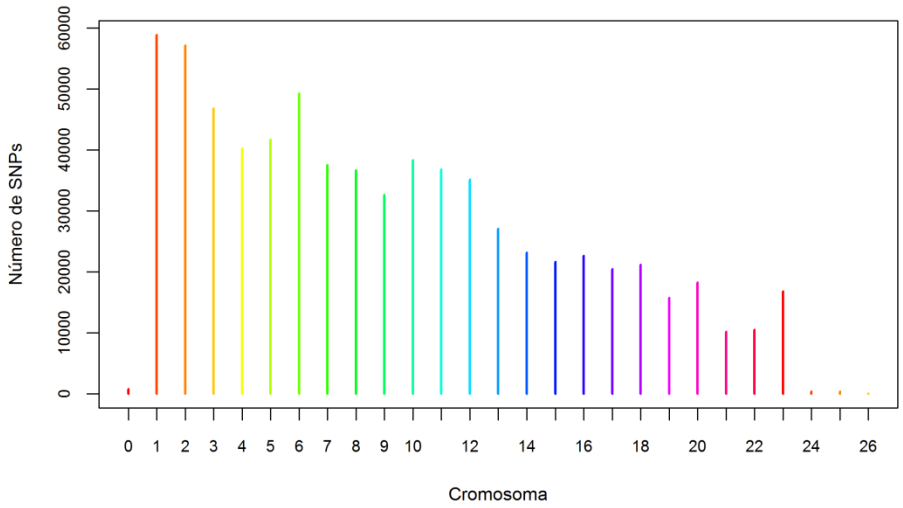


Figura - Anexos D Densidad de SNPs en la base de datos GWAS limpia. Se observa el número de SNPs por cada cromosoma una vez realizados los pasos de control de calidad por SNPs (n = 720815). Figura original generada en el programa *R statistics* (76).

Estratificación poblacional y estimación de la ancestría global

La reducción en la redundancia de SNPs *(o "pruning")* se realizó por separado para cada base de datos (CG o GWAS), previo a los pasos de evaluación de posible estratificación poblacional y de estimación de la ancestría global, según las recomendaciones descritas anteriormente (117). Todas las bases de datos a usar fueron actualizadas a las denominaciones del *"human genome build 37"*, en caso de ser necesario.

De la base de datos CG limpia con 1237 marcadores (ver **Figura - Anexos B**), se eliminó un SNP de cada par en desequilibrio de ligamiento (LD) con $R^2 > 0.2$, en ventanas de 50 SNPs y con cambios de ventana cada 5 SNPs. Debido a la baja densidad de SNPs en ésta base, no se consiguió un sobrelapamiento suficiente con las poblaciones de referencia amerindias del *HGDP* (114). Por lo anterior, se usaron poblaciones de referencia de la base de datos pública *HapMap3 project* (118) que incluyeron europeos [n = 112; CEU = residentes de Utah con ancestría europea], africanos [n = 90; LWK = africanos de Webuye, Kenia] y asiáticos [n = 84; CHB = asiáticos de Beijing, China]; éste último, teniendo en cuenta la similitud en frecuencias alélicas entre asiáticos y amerindios (121, 122). Adicionalmente y debido a la poca densidad de SNPs, se incluyó otra población mezclada como control [n = 50; MEX = residentes de California descendientes de mexicanos]. Con las poblaciones de referencia, se construyó una base de datos

consolidada y reducida que incluyó 473 SNPs sobrelapados (ver **Figura - Anexos E**), con el fin de evaluar la estratificación poblacional y estimar la ancestría global en 483 muestras colombianas (ver **Tabla - Anexos A**).

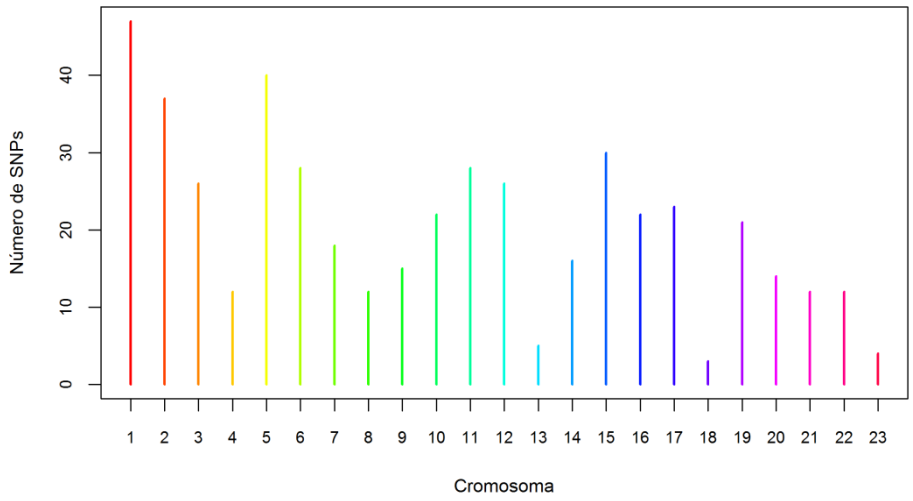


Figura - Anexos E Densidad de SNPs en la base de datos CG consolidada, con las poblaciones de referencia del *HapMap3 project*, y reducida. Se observa el número de SNPs por cada cromosoma una vez realizados los pasos de reducción en la redundancia de SNPs y la construcción de una base consolidada con las poblaciones de referencia (n = 473 SNPs). Figura original generada en el programa *R statistics* (76).

De la base de datos GWAS limpia con 720815 marcadores (ver **Figura - Anexos D**), se eliminó un SNP de cada par en LD con $R^2 > 0.1$, en ventanas de 50 SNPs y con cambios de ventana cada 10 SNPs. Se incluyeron poblaciones de referencia de las bases de datos públicas de *1000 Genomes* (113) y *HGDP* (114); debido a que las bases del *HGDP* (114) no tienen control de calidad por SNPs, excluimos aquellos con una tasa de error en el genotipado > 0.05 o con un MAF < 0.01, antes de usarlas. De la base de datos de *1000 Genomes* (113) se incluyeron las poblaciones europeas [n = 107; IBS = población de España] y africanas [n = 108; YRI = africanos de Ibadan, Nigeria], mientras que de la base de datos del *HGDP* (114) se incluyeron las poblaciones amerindias [n = 108; poblaciones de nativos americanos, AME= Pima, Maya, Karitiana, Surui y de Colombia]. Gracias a la alta densidad de SNPs, se logró construir una base de datos consolidada y reducida que incluyó 9868 SNPs sobrelapados con las poblaciones de referencia (ver **Figura - Anexos F**), con el fin de evaluar la estratificación poblacional y estimar la ancestría global en 415 muestras (ver **Tabla - Anexos A**).

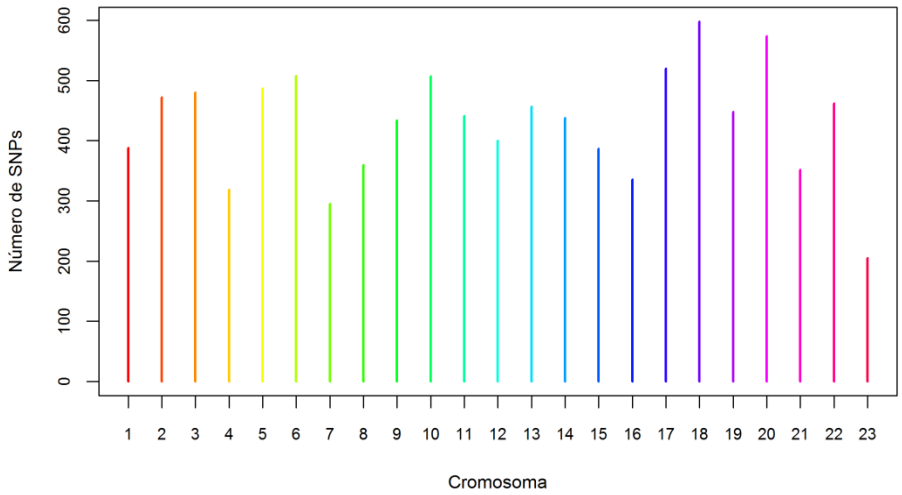


Figura - Anexos F Densidad de SNPs en la base de datos GWAS consolidada, con las poblaciones de referencia de *1000 Genomes* + *HGDP*, y reducida. Se observa el número de SNPs por cada cromosoma una vez realizados los pasos de reducción en la redundancia de SNPs y la construcción de una base consolidada con las poblaciones de referencia (n = 9868 SNPs). Figura original generada en el programa *R statistics* (76).

La evaluación de la estratificación poblacional se realizó en PLINK (119) mediante un análisis de escalamiento multidimensional (MDS), con las bases reducidas CG (ver **Figura - Anexos E**) y GWAS (ver **Figura - Anexos F**), por separado. Este análisis permitió ver las diferencias o similitudes de las muestras de Colombia (casos y controles) en relación con las poblaciones de referencia seleccionadas, al graficar las distancias en dos dimensiones (DIM). La estimación de la ancestría global de las muestras colombianas también se realizó por separado para las dos bases de datos reducidas, mediante un análisis no supervisado usando el algoritmo de ADMIXTURE v1.3.0 (115) y fijando el número de componentes ancestrales a k = 3. El MDS y los análisis de ancestría global se realizaron usando los genotipos en autosomas obtenidos para las muestras genotipadas con las dos plataformas, CG y GWAS (ver **Tabla - Anexos A**).

Adicionalmente y con el fin de evaluar sesgos de género, se realizaron análisis de ancestría en tres escenarios: en genoma completo (autosomas + cromosoma X), en autosomas y en cromosoma X; lo anterior, solo con los 415 individuos genotipados con la plataforma GWAS y usando la base de datos reducida (ver **Tabla - Anexos A** y **Figura - Anexos F**). Debido al carácter hemicigoto de los hombres con respecto al cromosoma X, siempre que se incluyeron marcadores en éste cromosoma se añadió el comando --haploid="male:23" (116). Una vez obtenidas las estimaciones en todas las 415 muestras,

se seleccionaron solo los *"controles"* de Colombia (n = 131; mujeres 73 y hombres 58), para hacer las respectivas comparaciones.

Estimación de la ancestría local a partir de datos de genoma completo (GWAS)

Para inferir la ancestría local (LAI) se usó el programa RFMix v1.5.4 (120), que se basa en un análisis iterativo. Inicia con el uso de genotipos de referencia de poblaciones ancestrales no mezcladas (o relativamente puras) para posteriormente involucrar la información genética de las poblaciones mezcladas en estudio (120). Lo anterior, con el fin de refinar el conocimiento sobre los patrones haplotípicos de las poblaciones ancestrales de referencia, mejorando así la precisión en la inferencia de la ancestría local en las muestras a estudio a través de múltiples pasos de un **algoritmo tipo *esperanza-maximización* (EM)** (120).

Éstos análisis LAI, se realizaron solo en las 415 muestras colombianas genotipadas con la plataforma GWAS (ver **Tabla - Anexos A**), por tener la densidad necesaria para este fin (720815 SNPs) (ver **Figura - Anexos D**). Como poblaciones ancestrales de referencia se incluyeron las mismas poblaciones de las bases de datos de *1000 Genomes* (113) y *HGDP* (114), usadas en las estimaciones de ancestría globales. Debido a que RFMix requiere que la información genética esté **faseada *(o en haplotipos)*,** se usó el programa SHAPEIT v2.778 (247) para estimar los haplotipos a partir de los genotipos de las muestras de Colombia y de las poblaciones de referencia del *HGDP* (114); para esto, se usó como poblaciones de referencia los genomas ya faseados de la base de datos de *1000 Genomes* (113). Una vez faseadas todas las bases de datos, éstas se fusionaron antes de correr los análisis en RFMix (120); la base de datos consolidada incluyó genotipos de 275284 SNPs (ver **Figura - Anexos G**). Se realizaron 5 iteraciones EM, como recomiendan los desarrolladores del programa (120), con el fin de corregir los análisis por el uso de poblaciones de referencia no tan puras, como las amerindias.

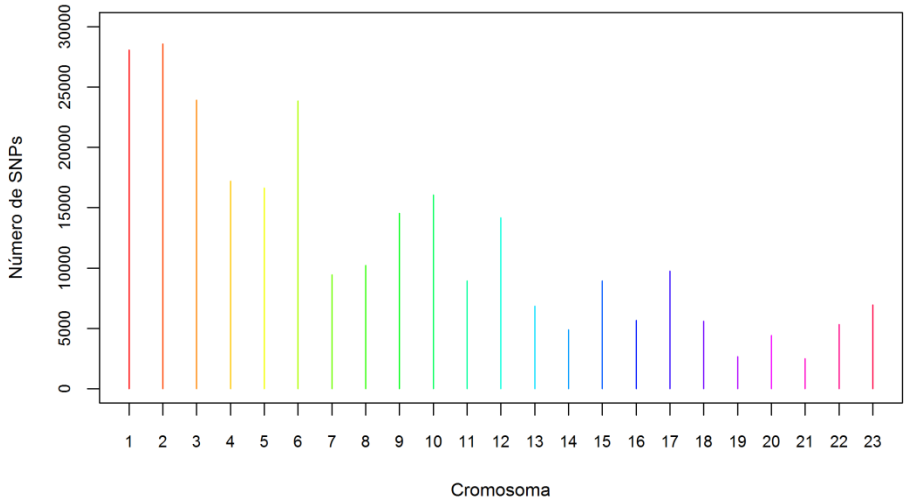


Figura - Anexos G Densidad de SNPs en la base de datos GWAS consolidada, con las poblaciones de referencia de *1000 Genomes* + *HGDP*, para hacer las estimaciones de ancestría local (LAI). Se observa el número de SNPs por cada cromosoma una vez preparada la base de datos consolidada con las poblaciones de referencia para RFMix (n = 275284 SNPs). Figura original generada en el programa *R statistics* (76).

Análisis estadístico

Los análisis para evaluar sesgos de género en 131 controles del estudio, genotipados con GWAS (ver **Tabla - Anexos A**) y con estimaciones de ancestría en genoma completo (autosomas + cromosoma X), en autosomas y en cromosoma X, se realizaron comparando las proporciones en autosomas versus cromosoma X con el test no paramétrico Wilcoxon-Mann-Whitney y un test basado en 100 mil permutaciones (116). Por otro lado, la magnitud de las diferencias en las ancestrías calculadas en el cromosoma X versus las calculadas en autosomas se obtuvo de acuerdo a la fórmula planteada por *Rishishwar L, et al* (9); la cual propone normalizar las diferencias de éstas proporciones con las ancestrías a nivel de genoma completo, para obtener la razón normalizada de éstas diferencias. Se calculó la proporción de hombres y mujeres de cada ancestría (europea, amerindia y africana), que participaron en el proceso de mestizaje de las ciudades y regiones de Colombia incluidas. Esto se realizó en base a las ecuaciones (a) y (b) descritas por *Goldberg A, et al* (87), que hablan de la contribución de hombres y mujeres en las medias observadas para cada ancestría en una población mezclada, asumiendo que ésta se compone de igual número de hombres que de mujeres y que resultó de un evento de mestizaje en un momento en el tiempo. Entonces, teniendo en cuenta que se estimaron las medias de cada ancestría a partir de las proporciones

observadas en autosomas ($H^{A}_{1,g,\delta} = Auto_{anc}$) y cromosoma X ($H^{X}_{1,g,\delta} = CrX_{anc}$) de los *"controles"* colombianos seleccionados, éstas se usaron para reemplazar los respectivos valores en las ecuaciones (a) y (b), que al desarrollarlas para despejar el porcentaje relativo de mujeres ($\%_{relativo,muj}$) y el porcentaje relativo de hombres ($\%_{relativo,hom}$), obtenemos que: el $\%_{relativo,muj} = 3CrX_{anc} - 2Auto_{anc}$ y el $\%_{relativo,hom} = 2Auto_{anc} - \%_{relativo,muj}$. Por último, el porcentaje de mujeres ($\%_{muj}$) se calculó así: $\%_{muj} = \%_{relativo,muj} / (\%_{relativo,muj} + \%_{relativo,hom})$; mientras que el porcentaje de hombres es: $\%_{hom} = (\%_{relativo,muj} + \%_{relativo,hom}) - \%_{muj}$.

Se realizó un análisis descriptivo basado en las características de 1019 individuos colombianos incluidos, con el fin de evaluar diferencias entre casos y controles, usando el estadístico χ^2 de Pearson.

Se obtuvieron estimaciones de la ancestría global en 813 muestras colombianas únicas y solo para estas fue posible ajustar los modelos de regresión logística por la ancestría. Se usó el coeficiente de correlación de Pearson (r) para evaluar la concordancia en las estimaciones de cada componente ancestral calculado con el programa ADMIXTURE (115) en 85 muestras repetidas entre las dos plataformas, CG y GWAS. Adicionalmente, se usó la ancestría local inferida con RFMix (120) para calcular la ancestría global en las muestras genotipadas con GWAS y correlacionar éstos estimados con los obtenidos con ADMIXTURE (115) en 415 muestras. Se usó la prueba de Wilcoxon-Mann-Whitney para evaluar las diferencias entre las proporciones de ancestrías de los casos (CCR y PA) en comparación con los controles.

Debido a que el tamaño de muestra del estudio es moderado y a que las proporciones de ancestría africana mostraron una distribución asimétrica hacia la derecha, se aplicó una transformación logit a las proporciones de ancestría, con el fin de lograr una distribución simétrica y evitar intervalos de confianza con la probabilidad de cobertura equivocada en los análisis de regresión ajustados. La ancestría amerindia se tomó como referencia y se incluyó la variable *"Array"* para corregir los análisis de regresiones logísticas por la plataforma o microarreglo usado en los cálculos de las ancestrías globales, CG o GWAS. Se corrieron modelos de regresión multinomial de los fenotipos PA y CCR, usando el grupo *"control"* como referencia, para evaluar su asociación con los componentes ancestrales ajustados por edad, sexo, nivel educativo, ciudad de origen, consumo de

AINES e historia familiar de CCR. Se escogió el mejor modelo según el criterio de información de Akaike (AIC).

Los análisis de asociación de genes candidatos se corrieron en las muestras genotipadas con la plataforma CG limpia, que incluye los genotipos de 1237 SNPs en 483 muestras colombianas (ver **Figura - Anexos B** y **Tabla - Anexos A**). Se realizó un análisis de asociación básico por SNP usando la prueba de χ^2 para evaluar diferencias en la frecuencia del alelo menor (MAF) entre casos y controles. Se corrieron regresiones logísticas de todos los 1237 SNPs ajustando por edad, ancestría europea, ancestría africana, y el *"Array"* (ya que para 85 muestras repetidas entre las dos plataformas, CG y GWAS, se usaron las ancestrías obtenidas con ésta última). Se tomó como valor significativo una $P < 0.05$ corregida por Bonferroni (equivalente a una $P < 4.04 \times 10^{-5}$). Siguiendo las recomendaciones descritas en el protocolo de *Clarke GM, et al* (145), se usó la prueba de permutación de valores de *P* basado en 100 mil replicados de un modelo alélico con el fin de identificar asociaciones significativas ($P < 0.05$).

Los análisis de asociación de genoma completo se realizaron en las muestras genotipadas con la plataforma GWAS limpia, incluyendo solo autosomas y cromosoma X (719571 SNPs en 415 muestras colombianas) (ver **Figura - Anexos D** y **Tabla - Anexos A**). Se realizó un análisis de asociación básico por SNP usando la prueba de χ^2, tomando como valor significativo una $P < 0.05$ corregida por Bonferroni (equivalente a una $P < 6.95 \times 10^{-8}$). Solo con el más significativo en los análisis por SNP, se corrieron las regresiones logísticas ajustadas por edad y ancestrías europea y africana (no se incluyó la variable *"Array"*, pues solo se usaron las ancestrías globales estimadas con GWAS); se tomó como significativo un valor de $P < 0.05$.

Lo que se espera en los estudios de asociación a gran escala, que incluye múltiples SNPs sin una hipótesis previa de asociación con la enfermedad de interés, es que solo unos pocos SNPs estén asociados con el fenotipo a estudio y que todos los demás se distribuyan bajo la hipótesis nula de no asociación. Sin embargo, confusores como son las diferencias alélicas por estratificación poblacional, parentesco entre las muestras o errores en el genotipado no detectados en los pasos de control de calidad por individuo y por SNPs, pueden inflar los resultados de SNPs asociados y resultar en falsos-positivos (248). Entonces, con el fin de corroborar que no existe estratificación poblacional en

nuestra muestra o sesgos de recolección o diferencias en el genotipado entre casos y controles, se calculó el factor de inflación lambda (λ) en los análisis de asociación y se observó la correlación entre el –log10(*P*) observado versus el esperado mediante un gráfico quantil-quantil. Adicionalmente, los resultados de los análisis de asociación se resumieron por cromosoma en un gráfico tipo Manhattan, que permite ver los –log10(*P*) obtenidos para cada SNP.

Con el fin de evaluar el papel de la variabilidad a nivel de genoma completo y de la ancestría local sobre el efecto de los SNPs encontradas como asociados, se corrieron modelos de regresión ajustados por los 10 primeros componentes principales (PCs) (ver **Figura - Anexos H**), junto con las proporciones de ancestría del cromosoma y del locus donde se ubica la variante seleccionada, además de las otras variantes como edad, sexo y ancestría global. Estos análisis se realizaron con las ~415 muestras que tienen información tipo GWAS (ver **Tabla - Anexos A**).

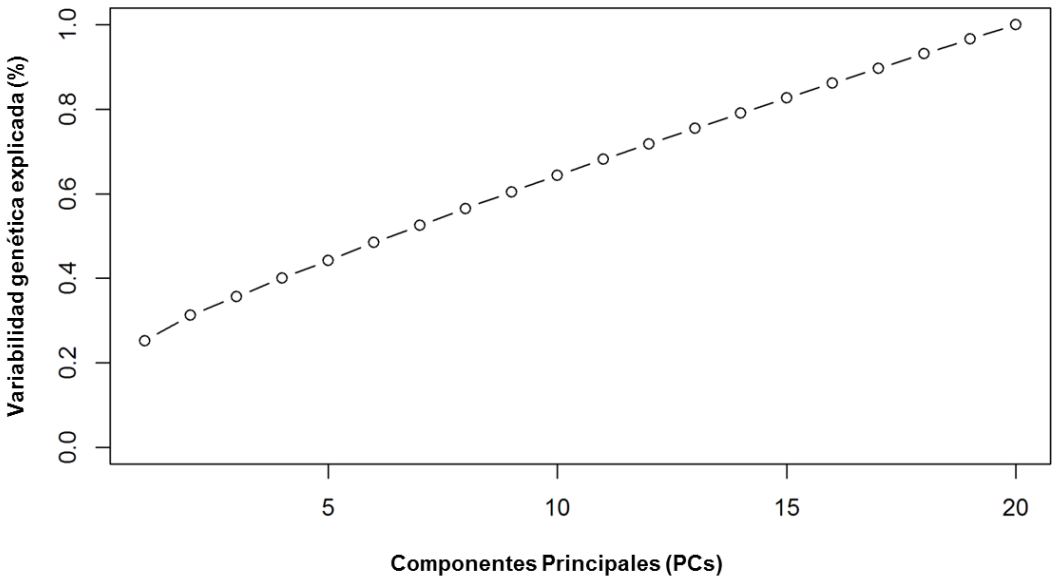


Figura - Anexos H Proporción de la variabilidad genómica explicada por 20 componentes principales (PCs). Se observa que con 10 PCs se explica el 64% de la variabilidad a nivel de genoma completo, según los análisis realizados en 415 muestras colombianas genotipadas con GWAS. Figura original generada en el programa *R statistics* (76).

Los análisis de componentes principales se corrieron en PLINK (119) y la sumatoria acumulada de la variabilidad genética que explican estos componentes se graficó en la **Figura - Anexos H**. Como se observa, los primeros 7 PCs explican un ~50% de la

variabilidad genética, mientras que los 10 primeros PCs explican un ~60%. Las proporciones de cada ancestría por cromosoma se obtuvieron a partir de los análisis tipo LAI, como el promedio de las ancestrías específicas de locus calculado con todos los marcadores incluidos a lo largo del cromosoma a estudio. De manera similar, se obtuvieron las proporciones de cada ancestría locus-específica, calculando el promedio de las ancestrías específicas de locus con los marcadores ubicados entre 100 mil pares de bases alrededor del SNP de interés.

Genotipado de los SNPs seleccionados en los análisis de asociación en colombianos y de los SNPs publicados a replicar

Alrededor de ~ 800 muestras del estudio fueron genotipadas para los SNPs 14q11.2:rs1760898 y 17q25.3:rs1065768, por mostrar asociación al riesgo de CCR y PA, respectivamente. Lo anterior se realizó mediante el kit TaqMan SNP Genotyping Assays (Applied Biosystems, Foster City, CA), según el protocolo. Los genotipos fueron asignados usando el Sequence Detection System (SDS) Software (Applied Biosystems, Foster City, CA). Para el SNP rs1760898 la correlación entre los genotipos obtenidos por CG versus Taqman fue de 99.15%, pues solo 4 de 469 fueron diferentes; igualmente, para el SNP rs1065768 la correlación entre los genotipos obtenidos por GWAS versus Taqman fue del 98.99%, ya que 4 de 399 fueron diferentes. Con el fin de confirmar las asociaciones observadas, se realizaron análisis básicos de asociación por SNP *(X^2)* y regresiones logísticas ajustadas por edad, sexo, ancestría europea, ancestría africana y el *"Array"*; se tomó como valor significativo una $P < 0.05$.

Se escogieron ~16 SNPs previamente publicados como asociados a CCR en poblaciones europeas hasta el 2011, para ser replicados en colombianos, usando ensayos de TaqMan. Se seleccionaron 14 SNPs que tuvieron una tasa de genotipado de 0.98. Se realizaron análisis básicos de asociación por SNP *(X^2)* y regresiones logísticas ajustadas por edad, sexo, ancestría europea, ancestría africana y el *"Array"*; se tomó como valor significativo una $P < 0.05$.

B. Anexo 2 – Materiales y métodos en análisis proteómicos

Selección de muestras

Para esta parte de la tesis con la cual se buscó explorar diferencias en los perfiles proteómicos entre casos de CCR y controles colombianos, se incluyeron muestras que habían sido captadas previamente para el estudio de marcadores de susceptibilidad genética *(ver Anexo 1 - Materiales y métodos en análisis genéticos)*. Por lo tanto, la inclusión de las muestras para estos análisis, estuvo condicionada por la autorización dada por los participantes en el consentimiento informado para el uso de sus muestras en estudios futuros. Adicionalmente, y con el fin de obtener la menor variabilidad biológica posible entre los individuos a analizar, se seleccionaron solo plasmas de participantes reclutados en condiciones de ayuno, de sexo masculino y con edades entre los 45 a 65 años. Teniendo en cuenta todo lo anterior, se incluyeron muestras de 20 casos de CCR y de 6 controles para los análisis proteómicos mediante cromatografía líquida en fase reversa acoplada a espectrometría de masas en tándem, RPLC/MS-MS.

Este trabajo en proteómica fue aprobado como un *"estudio sin riesgo"* por el Comité de Ética del Instituto Nacional de Cancerología de Bogotá, Colombia, según las guías para la investigación en humanos dispuestas en la Resolución 8430 de 1993 (De los Aspectos Éticos de la Investigación en Seres Humanos, Título II, Capítulo 1) publicado por el Ministerio de Salud, Republica de Colombia (https://www.invima.gov.co/images/pdf/medicamentos/resoluciones/etica_res_8430_1993.pdf). Lo anterior, teniendo en cuenta que se incluyeron muestras archivadas de un estudio previo.

Obtención de los plasmas libres de plaquetas

Una vez recolectadas las muestras de sangre periférica en tubos con el agente anticoagulante EDTA, estas se centrifugaron a 1300 x g por 10 minutos a 4°C para separar la capa de células blancas y el plasma. Ambos componentes sanguíneos se alicuotaron y almacenaron a -70°C. Posteriormente, las muestras de plasma seleccionadas para los análisis proteómicos se descongelaron en hielo y se centrifugaron nuevamente a 2400 x g por 15 minutos a 4°C, con el fin de obtener los plasmas libres de plaquetas, como recomienda la Organización del Proteoma Humano, HUPO (249), y se almacenaron a -70 °C hasta ser analizadas.

Reducción de la complejidad del plasma

En base a resultados previos del grupo de Investigación en Hormonas de la Universidad Nacional de Colombia, con respecto a dos métodos para reducir la complejidad del plasma para estudios proteómicos, se seleccionó el sistema ProteoPrep® Immunoaffinity Albumin and IgG Depletion, PROTIA (Sigma Aldrich, Missouri, US) porque permite la inmunodepleción de las dos proteínas más abundantes en el plasma mediante cromatografía de afinidad, ofreciendo un mejor control sobre las proteínas que se están descartando y por lo tanto una menor pérdida de información, lo que es importante en estudios proteómicos de tipo exploratorio (169). El proceso consiste en equilibrar la columna con 400 µL de una solución buffer de Tris pH 7.4 de baja fuerza iónica, seguido de centrifugación a 500 x g por 10 segundos, por 3 veces. Luego, 50 µL de plasma depletado en plaquetas se diluyeron en buffer de equilibrio hasta completar a 100µL y esta mezcla se aplicó a la columna a temperatura ambiente por 10 minutos de incubación. Un primer eluido inmunodepletado se obtuvo al centrifugar a 8000 x g por 60 segundos, el cual fue reaplicado en su totalidad a la columna, donde se dejó incubando por 10 minutos adicionales y se centrifugó nuevamente a 8000 x g por 60 segundos. Por último, se aplicaron 125 µL de buffer de equilibrio adicionales y se centrifugó nuevamente. Todo el eluido obtenido corresponde al plasma inmunodepletado y se almacenó a -20°C para posteriores análisis.

Digestión en gel

Todas las muestras depletadas se cuantificaron con el equipo NanoDrop 2000 (Thermo Scientific, Wilmington, DE) a una absorbancia UV A280 nm. De cada muestra se tomó el volumen necesario para diluir 25 µg de proteína en 10 µL del buffer carga NuPAGE LDS Sample Buffer, 4x (Invitrogen, Carlsbad, CA) y se completó a un total de 20 µL con 50 mM de bicarbonato de amonio, AMBIC. Las muestras se concentraron en condiciones denaturantes en un gel de poliacrilamida al 10%, Novex 10% Tris-Glycine Gel 1.0 m x 10 well (Invitrogen, Carlsbad, CA), usando el buffer de corrida 20X MOPS/SDS (Teknova, Hollister, CA), a 80 V – 400 mAMP – 10 W por 20 minutos. Los geles se tiñeron por 20 minutos con azul de Coomassie, Instant Blue Protein Stain (Expedeon, Harston, UK), para observar las proteínas y permitir su compatibilidad con la espectrometría de masas, MS. Para cada muestra se observó una banda gruesa de proteínas, la cual fue cortada en cuadros de aproximadamente 1 mm^3 que fueron almacenados a 4°C en tubos eppendorf.

La digestión en gel se realizó en simultáneo para todas las muestras usando tripsina, enzima serina-proteasa que corta las proteínas en el extremo carboxilo de aminoácidos lisina o arginina, con el fin de generar un patrón de digestión que dependerá de la secuencia de cada proteína. Este proceso incluyó los siguientes pasos: i) Cuatro lavados de los cortes con 200 µL de 50 mM AMBIC; ii) Deshidratación de los cortes con 200 µL de acetonitrilo, ACN, al 100%, por tres veces; iii) Reducción de los puentes disulfuro de las proteínas con la adición de 200 µL de 10 mM ditiotreitol, DTT, en 50 mM AMBIC, con la cual los cortes se dejaron incubando por 30 minutos a 56°C; iv) Nuevamente, se repitieron los pasos de deshidratación, con el fin de retirar el DTT; v) Alquilación o bloqueo de los grupos tiol de las proteínas reducidas con la adición de 250 µL de 55 mM iodoacetamida, IAA, en 50 mM AMBIC, con la cual los cortes se dejaron incubando por 20 min a temperatura ambiente y en la oscuridad; vi) Eliminación del exceso de IAA y dos lavados con 200 µL de 50 mM AMBIC; vii) Se repitieron los pasos de deshidratación, con el fin de retirar la IAA y se secaron los cortes en el SpeedVac SC110 (Thermo Scientific, Holbrook, NJ); viii) La digestión de las proteínas en gel se realizó mediante la incubación *overnight* a 37°C con 2 µL (1 µg/µL) de tripsina (Promega, Madison, WI, USA) en 200 µL de 50 mM AMBIC pH 8.0; ix) Al siguiente día, las muestras se centrifugaron y se recolectó el sobrenadante que contiene los péptidos en un tubo nuevo; x) Se añadió a los

cortes una solución de ACN al 60% con ácido trifluoroacético, TFA, al 0.1% hasta cubrirlos y se dejaron sonicando a baño maría por 10 minutos a temperatura ambiente; xi) Al terminar este último proceso de extracción de péptidos del gel, estos se centrifugaron por 10 minutos y se recolectaron con el sobrenadante, el cual fue adicionado a los péptidos colectados previamente; xii) Por último, los péptidos obtenidos se secaron en el SpeedVac hasta un volumen de 10 µL y fueron inmediatamente almacenados a -80°C hasta ser preparados para su análisis mediante RPLC/MS-MS.

RPLC/MS-MS

Se seleccionaron métodos en RPLC/MS-MS para la separación de los péptidos y la posterior identificación y cuantificación relativa de las proteínas, mediante conteo espectral o cuantificación sin marcaje.

Una vez los péptidos se secaron por completo, se reconstituyeron en 10 µL de una solución de ACN al 2% / TFA al 0.1% y se inyectaron en el equipo Easy-nLC II HPLC (Thermo Scientific, San Jose, CA) para su separación cromatográfica, el cual maneja volúmenes en el rango de nano litro por minuto (nL/min). Dentro del equipo HPLC, los péptidos primero pasaron por la columna Magic C18 100 Å, 5 Unit Reverse-Phase Trap (100 µm × 25 mm) para desalinizarlos, antes de ser separados de acuerdo a su hidrofobicidad en la columna Magic C18 200 Å, 3 Unit Reverse-Phase (75 µm × 150 mm). Los péptidos se eluyeron usando una fase móvil de 0.1% de ácido fórmico, FA (solvente A) y 100% de ACN (solvente B) con una tasa de flujo de 300 nL/min. Se usó un gradiente por 90 minutos variando el porcentaje del solvente B, así: 5% a 35% por 70 minutos, 35% a 80% por 8 minutos, 80% por 1 minuto, 80% a 5% por 1 minuto y 5% por 10 minutos, con lavados de 1 hora entre gradientes.

Los péptidos eluidos del sistema HPLC se ionizaron con la interfase Proxeon Nanospray Source (Thermo Scientific, San Jose, CA), que usa el método de ionización por ESI. Los iones se separaron y detectaron con el equipo Q-Exactive Orbitrap (Thermo Scientific, San Jose, CA) usando el modo dependiente de datos con un SCAN MS de iones precursores, seguido de 15 SCANs MS-MS. Se aplicó el parámetro de exclusión dinámica por 15 segundos; este parámetro permite evitar la detección repetitiva del ion

más abundante encontrado en un paquete de iones de similar m/z pulsados al analizador, puesto que una vez detectado, este es colocado en una lista de exclusión para ser ignorado por un tiempo determinado, con el fin de aumentar la probabilidad de detección de otros iones con similar m/z pero menos abundantes. La información del análisis MS se obtuvo con una resolución de 70,000, estableciendo como parámetro de control de pulsos de iones (AGC, Automatic Gain Control) un total de 1×10^6 iones o un tiempo de inyección de 2 ms. A su vez, la información del análisis MS-MS se obtuvo con una resolución de 17,500, estableciendo como parámetro AGC un total de 5×10^4 iones o un tiempo de inyección de 60 ms. La fragmentación de los péptidos se realizó mediante un método de disociación por colisión de alta energía, HCD, con una energía de colisión normalizada de 27. Se excluyeron de la fragmentación MS-MS los iones sin carga asignada o iones +1 y > +5.

Búsqueda en base de datos

Los espectros de masas se analizaron usando X! Tandem (The Global Proteome Machine, http://www.thegpm.org, fecha de acceso el 13 de febrero del 2014, versión CYCLONE 2013.02.01.1). X! Tandem fue configurado para buscar en la base de datos Uniprot Human Reference, especificando que se usó tripsina para generar los péptidos. Los parámetros de búsqueda fueron: carbamidometilo de cisteína como modificación fija, una tolerancia de la masa de los fragmentos de iones de 20 ppm y una tolerancia de la masa del ion precursor de 20 ppm. También se estableció como modificaciones variables las siguientes: glutamato → piro-glutamato en el N-terminal, pérdida de amonio en el N-terminal, glutamina → piro-glutamina en el N-terminal, desaminación de asparagina y glutamina, oxidación de metionina y triptófano, dioxidación de metionina y triptófano, y acetilación en el N-terminal.

Criterios para la identificación de proteínas

Para validar la identificación de péptidos y proteínas, se usó el programa Scaffold v4.4.1 (Proteome Software, Portland, OR). Se aceptaron identificaciones de péptidos con una probabilidad ≥ 99% (False Discovery Rate, FDR de 0.6%) e identificaciones de proteínas con una probabilidad ≥ 95%, con mínimo 2 péptidos (FDR de 0.02%). El programa usa el

algoritmo Protein Prophet (250) para calcular las probabilidades en la identificación de las proteínas. Se obtuvo un total de 469 proteínas y 437,438 péptidos usando los anteriores filtros.

Análisis estadístico

Los análisis estadísticos para comparar entre casos de CCR y controles se realizaron usando el paquete DESeq2 (251) en R statistic (76), el cual usa datos crudos de conteos y los normaliza según la función a usar. Los conteos espectrales totales sin normalizar se exportaron desde Scaffold.

Para realizar el análisis de componentes principales, PCA, se usaron los conteos transformados a una escala logarítmica en base 2, log2, normalizados con respecto al tamaño de la librería, y se graficaron los dos primeros PCs que capturan la mayor proporción de la variabilidad en la expresión de proteínas, con el fin de evaluar si existe una adecuada separación de los grupos (251).

Los análisis de expresión diferencial de proteínas entre casos de CCR y controles, se realizaron con la función DESeq, la cual primero estima el tamaño de los factores, así: i) calcula la media geométrica de los conteos de cada proteína entre todas las muestras; ii) estima la profundidad de la secuenciación de cada muestra en relación a todas las muestras, calculando para cada proteína el cociente de los conteos en cada muestra dividido por los conteos en todas las muestras; y iii) toma la mediana de todos los cocientes para estimar la profundidad relativa de la librería (251). Como segundo paso, la función DESeq también estima los valores de dispersión para cada proteína y finalmente, corre un modelo lineal generalizado, GLM, el cual incluye el tamaño de los factores y la dispersión estimada para cada proteína (251). A partir de los resultados, se seleccionaron las proteínas con diferencias significativas entre los grupos usando como significativo un valor de *P* corregido por FDR < 0.05. Adicionalmente, se aplicó un segundo filtro en base a un *fold change* ≥ 2.0 (el doble de expresión) o ≤ 0.5 (expresión a la mitad), con el fin de seleccionar las proteínas con mayores diferencias en su expresión y evaluar su capacidad de clasificación jerárquica de los grupos mediante un análisis no supervisado tipo HeatMap de los conteos transformados a una escala log2 y

normalizados. Adicionalmente, y con el propósito de confirmar la capacidad de clusterización de las proteínas candidatas en un análisis probabilístico no supervisado, se usó el paquete Rmixmod (252) en R statistics (76), el cual usa un algoritmo tipo EM.

Bibliografía

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013 [Available from: http://globocan.iarc.fr.
2. Piñeros M, Pardo C, Gamboa O, Hernández G. Atlas de mortalidad por cáncer en Colombia. Instituto Nacional de Cancerología, Instituto Geográfico Agustín Codazzi. Bogotá: Imprenta Nacional de Colombia; 2010.
3. Bedoya G, Montoya P, Garcia J, Soto I, Bourgeois S, Carvajal L, et al. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(19):7234-9.
4. CÓRDOBA L, GARCÍA JJ, HOYOS LS, DUQUE C, ROJAS W, CARVAJAL S, et al. COMPOSICIÓN GENÉTICA DE UNA POBLACIÓN DEL SUROCCIDENTE DE COLOMBIA. Revista Colombiana de Antropología. 2012;48:21-48.
5. Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet. 2012;8(3):e1002554.
6. Ibarra A, Restrepo T, Rojas W, Castillo A, Amorim A, Martinez B, et al. Evaluating the X chromosome-specific diversity of Colombian populations using insertion/deletion polymorphisms. PloS one. 2014;9(1):e87202.
7. Ossa H, Aquino J, Sierra S, Ramírez A, Carvalho EF, Gusmão L. Analysis of admixture in Native American populations from Colombia. Forensic Science International: Genetics Supplement Series. 2015;5:e332-e4.
8. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, et al. A genomewide admixture map for Latino populations. American journal of human genetics. 2007;80(6):1024-36.
9. Rishishwar L, Conley AB, Wigington CH, Wang L, Valderrama-Aguirre A, Jordan IK. Ancestry, admixture and fitness in Colombian genomes. Scientific reports. 2015;5:12376.
10. Rojas W, Parra MV, Campo O, Caro MA, Lopera JG, Arias W, et al. Genetic make up and structure of Colombian populations by means of uniparental and biparental DNA markers. Am J Phys Anthropol. 2010;143(1):13-20.
11. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, et al. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. 2008;4(3):e1000037.
12. Xavier C, Builes JJ, Gomes V, Ospino JM, Aquino J, Parson W, et al. Admixture and genetic diversity distribution patterns of non-recombining lineages of Native American ancestry in Colombian populations. PloS one. 2015;10(3):e0120155.
13. Winawer SJ. Natural history of colorectal cancer. The American journal of medicine. 1999;106(1A):3S-6S; discussion 50S-1S.

14. de la Chapelle A. Genetic predisposition to colorectal cancer. Nat Rev Cancer. 2004;4(10):769-80.
15. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. Nature reviews Genetics. 2009;10(6):353-8.
16. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
17. Sieber OM, Lamlum H, Crabtree MD, Rowan AJ, Barclay E, Lipton L, et al. Whole-gene APC deletions cause classical familial adenomatous polyposis, but not attenuated polyposis or "multiple" colorectal adenomas. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(5):2954-8.
18. Lindor NM. Hereditary colorectal cancer: MYH-associated polyposis and other newly identified disorders. Best practice & research Clinical gastroenterology. 2009;23(1):75-87.
19. Sieber OM, Lipton L, Crabtree M, Heinimann K, Fidalgo P, Phillips RK, et al. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. The New England journal of medicine. 2003;348(9):791-9.
20. Chhibber V, Dresser K, Mahalingam M. MSH-6: extending the reliability of immunohistochemistry as a screening tool in Muir-Torre syndrome. Mod Pathol. 2007;21(2):159-64.
21. Gallione C, Aylsworth AS, Beis J, Berk T, Bernhardt B, Clark RD, et al. Overlapping spectra of SMAD4 mutations in juvenile polyposis (JP) and JP-HHT syndrome. American journal of medical genetics Part A. 2010;152A(2):333-9.
22. Langeveld D, van Hattem WA, de Leng WW, Morsink FH, Ten Kate FJ, Giardiello FM, et al. SMAD4 immunohistochemistry reflects genetic status in juvenile polyposis syndrome. Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(16):4126-34.
23. O'Riordan JM, O'Donoghue D, Green A, Keegan D, Hawkes LA, Payne SJ, et al. Hereditary mixed polyposis syndrome due to a BMPR1A mutation. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2010;12(6):570-3.
24. Papp J, Kovacs ME, Solyom S, Kasler M, Borresen-Dale AL, Olah E. High prevalence of germline STK11 mutations in Hungarian Peutz-Jeghers Syndrome patients. BMC medical genetics. 2010;11:169.
25. Li XL, Zhou J, Chen ZR, Chng WJ. P53 mutations in colorectal cancer - molecular pathogenesis and pharmacological reactivation. World J Gastroenterol. 2015;21(1):84-93.
26. Teresi RE, Zbuk KM, Pezzolesi MG, Waite KA, Eng C. Cowden syndrome-affected patients with PTEN promoter mutations demonstrate abnormal protein translation. American journal of human genetics. 2007;81(4):756-67.
27. Institute: NC. PDQ® Genetics of Colorectal Cancer. Bethesda, MD: National Cancer Institute. ; [updated February 12, 2016. Available from: http://www.cancer.gov/types/colorectal/hp/colorectal-genetics-pdq.
28. Moran A, Ortega P, de Juan C, Fernandez-Marcelo T, Frias C, Sanchez-Pernaute A, et al. Differential colorectal carcinogenesis: Molecular basis and clinical relevance. World journal of gastrointestinal oncology. 2010;2(3):151-8.

29. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic alterations during colorectal-tumor development. The New England journal of medicine. 1988;319(9):525-32.
30. Azzoni C, Bottarelli L, Campanini N, Di Cola G, Bader G, Mazzeo A, et al. Distinct molecular patterns based on proximal and distal sporadic colorectal cancer: arguments for different mechanisms in the tumorigenesis. International journal of colorectal disease. 2007;22(2):115-26.
31. Delattre O, Olschwang S, Law DJ, Melot T, Remvikos Y, Salmon RJ, et al. Multiple genetic alterations in distal and proximal colorectal cancer. Lancet. 1989;2(8659):353-6.
32. Reichmann A, Levin B, Martin P. Human large-bowel cancer: correlation of clinical and histopathological features with banded chromosomes. International journal of cancer. 1982;29(6):625-9.
33. Nambiar PR, Gupta RR, Misra V. An "Omics" based survey of human colon cancer. Mutation research. 2010;693(1-2):3-18.
34. Al-Tassan NA, Whiffin N, Hosking FJ, Palles C, Farrington SM, Dobbins SE, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Scientific reports. 2015;5:10442.
35. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nature genetics. 2007;39(11):1315-7.
36. Cui R, Okada Y, Jang SG, Ku JL, Park JG, Kamatani Y, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut. 2011;60(6):799-805.
37. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nature genetics. 2012;44(7):770-6.
38. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nature genetics. 2010;42(11):973-7.
39. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, Broderick P, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nature genetics. 2008;40(1):26-8.
40. Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, Jee SH, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nature genetics. 2013;45(2):191-6.
41. Jiang K, Sun Y, Wang C, Ji J, Li Y, Ye Y, et al. Genome-wide association study identifies two new susceptibility loci for colorectal cancer at 5q23.3 and 17q12 in Han Chinese. Oncotarget. 2015;6(37):40327-36.
42. Lemire M, Qu C, Loo LW, Zaidi SH, Wang H, Berndt SI, et al. A genome-wide association study for colorectal cancer identifies a risk locus in 14q23.1. Hum Genet. 2015;134(11-12):1249-62.
43. Peters U, Hutter CM, Hsu L, Schumacher FR, Conti DV, Carlson CS, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. Hum Genet. 2012;131(2):217-34.
44. Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, Baron JA, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013;144(4):799-807 e24.

45. Real LM, Ruiz A, Gayan J, Gonzalez-Perez A, Saez ME, Ramirez-Lorca R, et al. A colorectal cancer susceptibility new variant at 4q26 in the Spanish population identified by genome-wide association analysis. PloS one. 2014;9(6):e101178.
46. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Raskin L, Lejbkowicz F, et al. A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. Carcinogenesis. 2014;35(11):2512-9.
47. Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun. 2015;6:7138.
48. Study C, Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nature genetics. 2008;40(12):1426-35.
49. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nature genetics. 2008;40(5):631-7.
50. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nature genetics. 2007;39(8):984-8.
51. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, Tenesa A, Jones AM, Howarth K, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. PLoS Genet. 2011;7(6):e1002105.
52. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nature genetics. 2008;40(5):623-30.
53. Wang H, Burnett T, Kono S, Haiman CA, Iwasaki M, Wilkens LR, et al. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun. 2014;5:4613.
54. Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, et al. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. Hum Mol Genet. 2014;23(17):4729-37.
55. Zhang B, Jia WH, Matsuda K, Kweon SS, Matsuo K, Xiang YB, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nature genetics. 2014;46(6):533-42.
56. Zhang B, Jia WH, Matsuo K, Shin A, Xiang YB, Matsuda K, et al. Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. International journal of cancer. 2014;135(4):948-55.
57. Bertucci F, Salas S, Eysteries S, Nasser V, Finetti P, Ginestier C, et al. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene. 2004;23(7):1377-91.
58. Carrer A, Zacchigna S, Balani A, Pistan V, Adami A, Porcelli F, et al. Expression profiling of angiogenic genes for the characterisation of colorectal carcinoma. European journal of cancer. 2008;44(12):1761-9.
59. Galamb O, Sipos F, Solymosi N, Spisak S, Krenacs T, Toth K, et al. Diagnostic mRNA expression patterns of inflamed, benign, and malignant colorectal biopsy specimen and their

correlation with peripheral blood results. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(10):2835-45.
60. Glebov OK, Rodriguez LM, Soballe P, DeNobile J, Cliatt J, Nakahara K, et al. Gene expression patterns distinguish colonoscopically isolated human aberrant crypt foci from normal colonic mucosa. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2006;15(11):2253-62.
61. Kim IJ, Kang HC, Jang SG, Kim K, Ahn SA, Yoon HJ, et al. Oligonucleotide microarray analysis of distinct gene expression patterns in colorectal cancer tissues harboring BRAF and K-ras mutations. Carcinogenesis. 2006;27(3):392-404.
62. Dowling P, Hughes DJ, Larkin AM, Meiller J, Henry M, Meleady P, et al. Elevated levels of 14-3-3 proteins, serotonin, gamma enolase and pyruvate kinase identified in clinical samples from patients diagnosed with colorectal cancer. Clinica chimica acta; international journal of clinical chemistry. 2015;441:133-41.
63. Fan NJ, Gao JL, Liu Y, Song W, Zhang ZY, Gao CF. Label-free quantitative mass spectrometry reveals a panel of differentially expressed proteins in colorectal cancer. BioMed research international. 2015;2015:365068.
64. Han M, Liew CT, Zhang HW, Chao S, Zheng R, Yip KT, et al. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2008;14(2):455-60.
65. Mazzanti R, Solazzo M, Fantappie O, Elfering S, Pantaleo P, Bechi P, et al. Differential expression proteomics of human colon cancer. American journal of physiology Gastrointestinal and liver physiology. 2006;290(6):G1329-38.
66. Ibrahim AE, Arends MJ, Silva AL, Wyllie AH, Greger L, Ito Y, et al. Sequential DNA methylation changes are associated with DNMT3B overexpression in colorectal neoplastic progression. Gut. 2011;60(4):499-508.
67. Vaughn CP, Wilson AR, Samowitz WS. Quantitative evaluation of CpG island methylation in hyperplastic polyps. Mod Pathol. 2010;23(1):151-6.
68. Cavalli-Sforza L, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press; 1994.
69. Jobling M, Hollox E, Kivisild T, Tyler-Smith C. Human Evolutionary Genetics. Second, editor: Garland Science; 2013.
70. Carvajal-Carmona LG, Soto ID, Pineda N, Ortiz-Barrientos D, Duque C, Ospina-Duque J, et al. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. American journal of human genetics. 2000;67(5):1287-95.
71. Bushnell D. The Making of Modern Colombia: A Nation in Spite of Itself. California: University of California Press; 1993.
72. Hudson R. Race and Ethnicity: Indigenous Peoples. In: ed NgpEt, editor. “Colombia: A Country Study” area handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress. ; 2010. p. 82-6.
73. Hudson R. Historical Setting: Early Colombia and Race and Ethnicity: Indigenous Peoples. In: ed NgpEt, editor. “Colombia: A Country Study”. area handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 4-6; 82-6.

74. Hudson R. Historical Setting: The Spanish Conquest and Colonial Society. In: ed NgpEt, editor. "Colombia: A Country Study". a handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 6-7.
75. Hudson R. Historical Setting: Colonial Society and Economy. In: ed NgpEt, editor. "Colombia: A Country Study. a handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 8-11.
76. Team TRc. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Viena, Austria2014.
77. Hudson R. Historical Setting: Colonial Society and Economy and Race and Ethnicity: Racial Distinctions. In: ed NgpEt, editor. "Colombia: A Country Study" a handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 8-11; 86-90.
78. Hudson R. Race and Ethnicity: Indigenous Peoples and Race and Ethnicity: Racial Distinctions. In: ed NgpEt, editor. "Colombia: A Country Study". a handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 82-90.
79. Departamento Nacional de Estadística D. La visibilización estadística de los grupos étnicos colombianos.2007.
80. Departamento Nacional de Estadística D. Colombia, una nación multicultural.2007.
81. Hudson R. Race and Ethnicity: Racial Distinctions. In: ed NgpEt, editor. "Colombia: A Country Study" a handbook series: a country study. Washington, D.C.: Federal Research Division, Library of Congress.; 2010. p. 86-90.
82. Cidse., XXI. A, Departamento Administrativo Nacional de Estadística D. Cuantos somos Como vamos Diagnóstico Sociodemográfico de Cali y 10 municipios del Pacífico nariñense. Editado en Santiago de Cali, Colombia: Afroamérica XXI; 2011.
83. Yunis JJ, Yunis EJ. Mitochondrial DNA (mtDNA) haplogroups in 1526 unrelated individuals from 11 Departments of Colombia. Genet Mol Biol. 2013;36(3):329-35.
84. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. American journal of human genetics. 2015;96(1):37-53.
85. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proceedings of the National Academy of Sciences of the United States of America. 2010;107 Suppl 2:8954-61.
86. Emery LS, Felsenstein J, Akey JM. Estimators of the human effective sex ratio detect sex biases on different timescales. American journal of human genetics. 2010;87(6):848-56.
87. Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. Genetics. 2015;201(1):263-79.
88. Fejerman L, John EM, Huntsman S, Beckman K, Choudhry S, Perez-Stable E, et al. Genetic ancestry and risk of breast cancer among U.S. Latinas. Cancer Res. 2008;68(23):9723-8.
89. Fejerman L, Romieu I, John EM, Lazcano-Ponce E, Huntsman S, Beckman KB, et al. European ancestry is positively associated with breast cancer risk in Mexican women. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2010;19(4):1074-82.

90. Fejerman L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, et al. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Hum Mol Genet. 2012;21(8):1907-17.
91. Fejerman L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, et al. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat Commun. 2014;5:5260.
92. Fejerman L, Hu D, Huntsman S, John EM, Stern MC, Haiman CA, et al. Genetic ancestry and risk of mortality among U.S. Latinas with breast cancer. Cancer Res. 2013;73(24):7243-53.
93. Chlebowski RT, Chen Z, Anderson GL, Rohan T, Aragaki A, Lane D, et al. Ethnicity and breast cancer: factors influencing differences in incidence and outcome. Journal of the National Cancer Institute. 2005;97(6):439-48.
94. Jatoi I, Becher H, Leake CR. Widening disparity in survival between white and African-American patients with breast carcinoma treated in the U. S. Department of Defense Healthcare system. Cancer. 2003;98(5):894-9.
95. Newman LA, Griffith KA, Jatoi I, Simon MS, Crowe JP, Colditz GA. Meta-analysis of survival in African American and white American patients with breast cancer: ethnicity compared with socioeconomic status. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24(9):1342-9.
96. Fejerman L, Haiman CA, Reich D, Tandon A, Deo RC, John EM, et al. An admixture scan in 1,484 African American women with breast cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2009;18(11):3110-7.
97. Reding KW, Carlson CS, Kahsai O, Chen CC, McDavid A, Doody DR, et al. Examination of ancestral informative markers and self-reported race with tumor characteristics of breast cancer among Black and White women. Breast Cancer Res Treat. 2012;134(2):801-9.
98. Martin DN, Boersma BJ, Yi M, Reimers M, Howe TM, Yfantis HG, et al. Differences in the tumor microenvironment between African-American and European-American breast cancer patients. PloS one. 2009;4(2):e4531.
99. Bensen JT, Xu Z, McKeigue PM, Smith GJ, Fontham ET, Mohler JL, et al. Admixture mapping of prostate cancer in African Americans participating in the North Carolina-Louisiana Prostate Cancer Project (PCaP). Prostate. 2014;74(1):1-9.
100. Bock CH, Schwartz AG, Ruterbusch JJ, Levin AM, Neslund-Dudas C, Land SJ, et al. Results from a prostate cancer admixture mapping study in African-American men. Hum Genet. 2009;126(5):637-42.
101. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(38):14068-73.
102. Yang JJ, Cheng C, Devidas M, Cao X, Fan Y, Campana D, et al. Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nature genetics. 2011;43(3):237-41.
103. Ollberding NJ, Nomura AM, Wilkens LR, Henderson BE, Kolonel LN. Racial/ethnic differences in colorectal cancer risk: the multiethnic cohort study. International journal of cancer. 2011;129(8):1899-906.
104. Paquette IM, Ying J, Shah SA, Abbott DE, Ho SM. African Americans should be screened at an earlier age for colorectal cancer. Gastrointest Endosc. 2015;82(5):878-83.

105. Qaseem A, Denberg TD, Hopkins RH, Jr., Humphrey LL, Levine J, Sweet DE, et al. Screening for colorectal cancer: a guidance statement from the American College of Physicians. Ann Intern Med. 2012;156(5):378-86.
106. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. Am J Gastroenterol. 2009;104(3):739-50.
107. Cash BD, Banerjee S, Anderson MA, Ben-Menachem T, Decker GA, Fanelli RD, et al. Ethnic issues in endoscopy. Gastrointest Endosc. 2010;71(7):1108-12.
108. Stern MC, Fejerman L, Das R, Setiawan VW, Cruz-Correa MR, Perez-Stable EJ, et al. Variability in Cancer Risk and Outcomes Within US Latinos by National Origin and Genetic Ancestry. Curr Epidemiol Rep. 2016;3:181-90.
109. Pinheiro PS, Sherman RL, Trapido EJ, Fleming LE, Huang Y, Gomez-Marin O, et al. Cancer incidence in first generation U.S. Hispanics: Cubans, Mexicans, Puerto Ricans, and new Latinos. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2009;18(8):2162-9.
110. Stern MC, Zhang J, Lee E, Deapen D, Liu L. Disparities in colorectal cancer incidence among Latino subpopulations in California defined by country of origin. Cancer Causes Control. 2016;27(2):147-55.
111. Pardo C, Cendales R. Incidencia, mortalidad y prevalencia de cáncer en Colombia, 2007-2011. Primera Edición ed. Bogotá, DC.: Instituto Nacional de Cancerología; 2015. 148 p.
112. Pardo C, Cendales R. Incidencia estimada y mortalidad por cancer en Colombia, 2002 - 2006. Bogotá, DC: Instituto Nacional de Cancerología; 2010. 143 p.
113. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
114. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, et al. A human genome diversity cell line panel. Science. 2002;296(5566):261-2.
115. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655-64.
116. Shringarpure SS, Bustamante CD, Lange K, Alexander DH. Efficient analysis of large datasets and sex bias with ADMIXTURE. BMC Bioinformatics. 2016;17:218.
117. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73.
118. International HapMap C. The International HapMap Project. Nature. 2003;426(6968):789-96.
119. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.
120. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. American journal of human genetics. 2013;93(2):278-88.
121. Hernandez-Suarez G, Sanabria MC, Serrano M, Herran OF, Perez J, Plata JL, et al. Genetic ancestry is associated with colorectal adenomas and adenocarcinomas in Latino populations. Eur J Hum Genet. 2014;22(10):1208-16.

122. Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA, et al. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet. 2004;114(3):263-71.
123. Amersi F, Agustin M, Ko CY. Colorectal cancer: epidemiology, risk factors, and health services. Clin Colon Rectal Surg. 2005;18(3):133-40.
124. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. The New England journal of medicine. 2000;343(2):78-85.
125. Grady WM. Genetic testing for high-risk colon cancer patients. Gastroenterology. 2003;124(6):1574-94.
126. Friis S, Riis AH, Erichsen R, Baron JA, Sorensen HT. Low-Dose Aspirin or Nonsteroidal Anti-inflammatory Drug Use and Colorectal Cancer Risk: A Population-Based, Case-Control Study. Ann Intern Med. 2015;163(5):347-55.
127. Oshima M, Dinchuk JE, Kargman SL, Oshima H, Hancock B, Kwong E, et al. Suppression of intestinal polyposis in Apc delta716 knockout mice by inhibition of cyclooxygenase 2 (COX-2). Cell. 1996;87(5):803-9.
128. Stolfi C, Fina D, Caruso R, Caprioli F, Sarra M, Fantini MC, et al. Cyclooxygenase-2-dependent and -independent inhibition of proliferation of colon cancer cells by 5-aminosalicylic acid. Biochem Pharmacol. 2008;75(3):668-76.
129. Colotta F, Allavena P, Sica A, Garlanda C, Mantovani A. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. Carcinogenesis. 2009;30(7):1073-81.
130. Pardo-Seco J, Martinon-Torres F, Salas A. Evaluating the accuracy of AIM panels at quantifying genome ancestry. BMC Genomics. 2014;15:543.
131. Ma Y, Zhao J, Wong JS, Ma L, Li W, Fu G, et al. Accurate inference of local phased ancestry of modern admixed populations. Scientific reports. 2014;4:5800.
132. Williams RC, Elston RC, Kumar P, Knowler WC, Abboud HE, Adler S, et al. Selecting SNPs informative for African, American Indian and European Ancestry: application to the Family Investigation of Nephropathy and Diabetes (FIND). BMC Genomics. 2016;17:325.
133. Johnson NA, Coram MA, Shriver MD, Romieu I, Barsh GS, London SJ, et al. Ancestral components of admixed genomes in a Mexican cohort. PLoS Genet. 2011;7(12):e1002410.
134. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, et al. Reconstructing the population genetic history of the Caribbean. PLoS Genet. 2013;9(11):e1003925.
135. Kehdy FS, Gouveia MH, Machado M, Magalhaes WC, Horimoto AR, Horta BL, et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(28):8696-701.
136. de Moura RR, de Queiroz Balbino V, Crovella S, Brandao LA. On the use of Chinese population as a proxy of Amerindian ancestors in genetic admixture studies with Latin American populations. Eur J Hum Genet. 2016;24(3):326-7.
137. Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet. 2009;76(1):1-18.

138. Piepoli A, Santoro R, Cristofaro G, Traversa GP, Gennarelli M, Accadia L, et al. Linkage analysis identifies gene carriers among members of families with hereditary nonpolyposis colorectal cancer. Gastroenterology. 1996;110(5):1404-9.
139. Pulst SM. Genetic linkage analysis. Arch Neurol. 1999;56(6):667-72.
140. Manolio TA, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009;60:443-56.
141. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet. 2002;11(20):2417-23.
142. Lewis CM, Knight J. Introduction to genetic association studies. Cold Spring Harbor protocols. 2012;2012(3):297-306.
143. Schmit SL, Schumacher FR, Edlund CK, Conti DV, Ihenacho U, Wan P, et al. Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis. 2016;37(6):547-56.
144. Fejerman L, Stern MC, Ziv E, John EM, Torres-Mejia G, Hines LM, et al. Genetic ancestry modifies the association between genetic risk variants and breast cancer risk among Hispanic and non-Hispanic white women. Carcinogenesis. 2013;34(8):1787-93.
145. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT. Basic statistical analysis in genetic case-control studies. Nature protocols. 2011;6(2):121-33.
146. Foulkes A. Applied Statistical Genetics with R For Population-based Association Studies. 1 ed. New York: Springer-Verlag New York; 2009. XXIII, 252 p.
147. He J, Wilkens LR, Stram DO, Kolonel LN, Henderson BE, Wu AH, et al. Generalizability and epidemiologic characterization of eleven colorectal cancer GWAS hits in multiple populations. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2011;20(1):70-81.
148. Wang J, Carvajal-Carmona LG, Chu JH, Zauber AG, Collaborators APCT, Kubo M, et al. Germline variants and advanced colorectal adenomas: adenoma prevention with celecoxib trial genome-wide association study. Clinical cancer research : an official journal of the American Association for Cancer Research. 2013;19(23):6430-7.
149. von Holst S, Picelli S, Edler D, Lenander C, Dalen J, Hjern F, et al. Association studies on 11 published colorectal cancer risk loci. Br J Cancer. 2010;103(4):575-80.
150. Qin Q, Liu L, Zhong R, Zou L, Yin J, Zhu B, et al. The genetic variant on chromosome 10p14 is associated with risk of colorectal cancer: results from a case-control study and a meta-analysis. PloS one. 2013;8(5):e64310.
151. Kupfer SS, Anderson JR, Hooker S, Skol A, Kittles RA, Keku TO, et al. Genetic heterogeneity in colorectal cancer associations between African and European americans. Gastroenterology. 2010;139(5):1677-85, 85 e1-8.
152. Carvajal-Carmona LG, Zauber AG, Jones AM, Howarth K, Wang J, Cheng T, et al. Much of the genetic risk of colorectal cancer is likely to be mediated through susceptibility to adenomas. Gastroenterology. 2013;144(1):53-5.
153. Scholefield JH. ABC of colorectal cancer: screening. BMJ. 2000;321(7267):1004-6.
154. Bienz M, Clevers H. Linking colorectal cancer to Wnt signaling. Cell. 2000;103(2):311-20.
155. Kolligs FT, Bommer G, Goke B. Wnt/beta-catenin/tcf signaling: a critical pathway in gastrointestinal tumorigenesis. Digestion. 2002;66(3):131-44.

156. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics. 2009;41(8):882-4.
157. Harismendy O, Frazer KA. Elucidating the role of 8q24 in colorectal cancer. Nature genetics. 2009;41(8):868-9.
158. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nature genetics. 2009;41(8):885-90.
159. Lewis A, Freeman-Mills L, de la Calle-Mustienes E, Giraldez-Perez RM, Davis H, Jaeger E, et al. A polymorphic enhancer near GREM1 influences bowel cancer risk through differential CDX2 and TCF7L2 binding. Cell Rep. 2014;8(4):983-90.
160. Kosinski C, Li VS, Chan AS, Zhang J, Ho C, Tsui WY, et al. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(39):15418-23.
161. Jones AM, Beggs AD, Carvajal-Carmona L, Farrington S, Tenesa A, Walker M, et al. TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. Gut. 2012;61(2):248-54.
162. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. Stat Sci. 2009;24(4):561-73.
163. Biancolella M, Fortini BK, Tring S, Plummer SJ, Mendoza-Fandino GA, Hartiala J, et al. Identification and characterization of functional risk variants for colorectal cancer mapping to chromosome 11q23.1. Hum Mol Genet. 2014;23(8):2198-209.
164. Qin L, Liu Y, Wang Y, Wu G, Chen J, Ye W, et al. Computational Characterization of Osteoporosis Associated SNPs and Genes Identified by Genome-Wide Association Studies. PloS one. 2016;11(3):e0150070.
165. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561-3.
166. Honore B, Ostergaard M, Vorum H. Functional genomics studied by proteomics. Bioessays. 2004;26(8):901-15.
167. Sallam RM. Proteomics in cancer biomarkers discovery: challenges and applications. Dis Markers. 2015;2015:321370.
168. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, et al. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev. 1996;13:19-50.
169. Rodríguez R, Urrego W, Sanabria-Salas M, Sánchez-Gómez M, Umaña-Perez A. Implementación de una metodología para la separación de proteomas de plasma humano mediante electroforesis bidimensional. Rev Colomb Quim. 2015;44(3):30-8.
170. Ahmad Y, Sharma N. An Effective Method for the Analysis of Human Plasma Proteome using Two-dimensional Gel Electrophoresis. J Proteomics Bioinform. 2009;2: 495-9.
171. Mitulovic G, Mechtler K. HPLC techniques for proteomics analysis--a short overview of latest developments. Brief Funct Genomic Proteomic. 2006;5(4):249-60.
172. Mann M, Hendrickson RC, Pandey A. Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem. 2001;70:437-73.
173. Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, et al. Standard operating procedures for serum and plasma collection: early detection research network consensus

statement standard operating procedure integration working group. J Proteome Res. 2009;8(1):113-7.
174. Rabilloud T, Lelong C. Two-dimensional gel electrophoresis in proteomics: a tutorial. J Proteomics. 2011;74(10):1829-41.
175. Karpievitch YV, Polpitiya AD, Anderson GA, Smith RD, Dabney AR. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. Ann Appl Stat. 2010;4(4):1797-823.
176. Shi Y, Xiang R, Horvath C, Wilkins JA. The role of liquid chromatography in proteomics. J Chromatogr A. 2004;1053(1-2):27-36.
177. Abian J. The coupling of gas and liquid chromatography with mass spectrometry. J Mass Spectrom. 1999;34:157–68.
178. Wasinger VC, Zeng M, Yau Y. Current status and advances in quantitative proteomic mass spectrometry. Int J Proteomics. 2013;2013:180605.
179. Skoog D, Holler F, Crouch S. Principles of Instrumental Analysis. Sixth Edition ed: Brooks/Cole ©; 2007.
180. Gilany K, Moens L, Dewilde S. Mass spectrometry based proteomics in the life sciences: a review. Journal of Paramedical Sciences. 2010;1(1):53–78.
181. Ardrey RE. Liquid chromatography-mass spectrometry : an introduction. New York: J. Wiley; 2003. xviii, 276 p. p.
182. Tuli L, Ressom HW. LC-MS Based Detection of Differential Protein Expression. J Proteomics Bioinform. 2009;2:416-38.
183. Madeira P, Alves P, Borges C. Fourier Transform - Materials Analysis: InTech; 2012 May 23. 272 p.
184. Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: instrumentation, ion motion and applications. Mass Spectrom Rev. 2008;27(6):661-99.
185. Zubarev RA, Makarov A. Orbitrap mass spectrometry. Anal Chem. 2013;85(11):5288-96.
186. Kalli A, Smith GT, Sweredoski MJ, Hess S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. J Proteome Res. 2013;12(7):3071-86.
187. Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, et al. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics. 2011;10(9):M111 011015.
188. Ang CS, Phung J, Nice EC. The discovery and validation of colorectal cancer biomarkers. Biomed Chromatogr. 2011;25(1-2):82-99.
189. Ahn SB, Han DS, Bae JH, Byun TJ, Kim JP, Eun CS. The Miss Rate for Colorectal Adenoma Determined by Quality-Adjusted, Back-to-Back Colonoscopies. Gut Liver. 2012;6(1):64-70.
190. Rex DK, Cutler CS, Lemmel GT, Rahmani EY, Clark DW, Helper DJ, et al. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. Gastroenterology. 1997;112(1):24-8.
191. Ahlquist DA. Molecular detection of colorectal neoplasia. Gastroenterology. 2010;138(6):2127-39.
192. Huijbers A, Velstra B, Dekker TJ, Mesker WE, van der Burgt YE, Mertens BJ, et al. Proteomic serum biomarkers and their potential application in cancer screening programs. Int J Mol Sci. 2010;11(11):4175-93.

193. Tanaka T, Tanaka M, Tanaka T, Ishigamori R. Biomarkers for colorectal cancer. Int J Mol Sci. 2010;11(9):3209-25.
194. Gonzalez-Pons M, Cruz-Correa M. Colorectal Cancer Biomarkers: Where Are We Now? BioMed research international. 2015;2015:149014.
195. de Wit M, Fijneman RJ, Verheul HM, Meijer GA, Jimenez CR. Proteomics in colorectal cancer translational research: biomarker discovery for clinical applications. Clin Biochem. 2013;46(6):466-79.
196. Luo Y, Wang L, Wang J. Developing proteomics-based biomarkers for colorectal neoplasms for clinical practice: opportunities and challenges. Proteomics Clin Appl. 2013;7(1-2):30-41.
197. Geyer PE, Kulak NA, Pichler G, Holdt LM, Teupser D, Mann M. Plasma Proteome Profiling to Assess Human Health and Disease. Cell Syst. 2016;2(3):185-95.
198. Dayon L, Kussmann M. Proteomics of human plasma: A critical comparison of analytical workflows in terms of effort, throughput and outcome. eu pa open proteomics. 2013;1:8–16.
199. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013;113(4):2343-94.
200. Boja ES, Rodriguez H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. Proteomics. 2012;12(8):1093-110.
201. Nanjappa V, Thomas JK, Marimuthu A, Muthusamy B, Radhakrishnan A, Sharma R, et al. Plasma Proteome Database as a resource for proteomics research: 2014 update. Nucleic Acids Res. 2014;42(Database issue):D959-65.
202. Xie F, Smith RD, Shen Y. Advanced proteomic liquid chromatography. J Chromatogr A. 2012;1261:78-90.
203. Naba A, Clauser KR, Whittaker CA, Carr SA, Tanabe KK, Hynes RO. Extracellular matrix signatures of human primary metastatic colon cancers and their metastases to liver. BMC Cancer. 2014;14:518.
204. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol. 2012;196(4):395-406.
205. Hynes RO, Naba A. Overview of the matrisome--an inventory of extracellular matrix constituents and functions. Cold Spring Harb Perspect Biol. 2012;4(1):a004903.
206. Hamm A, Veeck J, Bektas N, Wild PJ, Hartmann A, Heindrichs U, et al. Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: a systematic expression analysis. BMC Cancer. 2008;8:25.
207. Paris S, Sesboue R, Delpech B, Chauzy C, Thiberville L, Martin JP, et al. Inhibition of tumor growth and metastatic spreading by overexpression of inter-alpha-trypsin inhibitor family chains. International journal of cancer. 2002;97(5):615-20.
208. Chong PK, Lee H, Zhou J, Liu SC, Loh MC, Wang TT, et al. ITIH3 is a potential biomarker for early detection of gastric cancer. J Proteome Res. 2010;9(7):3671-9.
209. Subbannayya Y, Mir SA, Renuse S, Manda SS, Pinto SM, Puttamallesh VN, et al. Identification of differentially expressed serum proteins in gastric adenocarcinoma. J Proteomics. 2015;127(Pt A):80-8.
210. van den Broek I, Sparidans RW, van Winden AW, Gast MC, van Dulken EJ, Schellens JH, et al. The absolute quantification of eight inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4)-derived peptides in serum from breast cancer patients. Proteomics Clin Appl. 2010;4(12):931-9.

211. Ivancic MM, Irving AA, Jonakin KG, Dove WF, Sussman MR. The concentrations of EGFR, LRG1, ITIH4, and F5 in serum correlate with the number of colonic adenomas in ApcPirc/+ rats. Cancer Prev Res (Phila). 2014;7(11):1160-9.
212. Peltier J, Roperch JP, Audebert S, Borg JP, Camoin L. Quantitative proteomic analysis exploring progression of colorectal cancer: Modulation of the serpin family. J Proteomics. 2016;148:139-48.
213. Venning FA, Wullkopf L, Erler JT. Targeting ECM Disrupts Cancer Progression. Front Oncol. 2015;5:224.
214. Dano K, Behrendt N, Hoyer-Hansen G, Johnsen M, Lund LR, Ploug M, et al. Plasminogen activation and cancer. Thromb Haemost. 2005;93(4):676-81.
215. Lund LR, Green KA, Stoop AA, Ploug M, Almholt K, Lilla J, et al. Plasminogen activation independent of uPA and tPA maintains wound healing in gene-deficient mice. EMBO J. 2006;25(12):2686-97.
216. da Costa PL, Sirois P, Tannock IF, Chammas R. The role of kinin receptors in cancer and therapeutic opportunities. Cancer Lett. 2014;345(1):27-38.
217. Matsumura Y, Maruo K, Kimura M, Yamamoto T, Konno T, Maeda H. Kinin-generating cascade in advanced cancer patients and in vitro study. Jpn J Cancer Res. 1991;82(6):732-41.
218. Perez-Holanda S, Blanco I, Menendez M, Rodrigo L. Serum concentration of alpha-1 antitrypsin is significantly higher in colorectal cancer patients than in healthy controls. BMC Cancer. 2014;14:355.
219. Bujanda L, Sarasqueta C, Cosme A, Hijona E, Enriquez-Navascues JM, Placer C, et al. Evaluation of alpha 1-antitrypsin and the levels of mRNA expression of matrix metalloproteinase 7, urokinase type plasminogen activator receptor and COX-2 for the diagnosis of colorectal cancer. PloS one. 2013;8(1):e51810.
220. El-Akawi ZJ, Al-Hindawi FK, Bashir NA. Alpha-1 antitrypsin (alpha1-AT) plasma levels in lung, prostate and breast cancer patients. Neuro Endocrinol Lett. 2008;29(4):482-4.
221. Cavalcante Mde S, Torres-Romero JC, Lobo MD, Moreno FB, Bezerra LP, Lima DS, et al. A panel of glycoproteins as candidate biomarkers for early diagnosis and treatment evaluation of B-cell acute lymphoblastic leukemia. Biomark Res. 2016;4:1.
222. Zelvyte I, Lindgren S, Janciauskiene S. Multiple effects of alpha1-antitrypsin on breast carcinoma MDA-MB 468 cell growth and invasiveness. Eur J Cancer Prev. 2003;12(2):117-24.
223. Zelvyte I, Stevens T, Westin U, Janciauskiene S. alpha1-antitrypsin and its C-terminal fragment attenuate effects of degranulated neutrophil-conditioned medium on lung cancer HCC cells, in vitro. Cancer Cell Int. 2004;4(1):7.
224. Sahni A, Simpson-Haidaris PJ, Sahni SK, Vaday GG, Francis CW. Fibrinogen synthesized by cancer cells augments the proliferative effect of fibroblast growth factor-2 (FGF-2). J Thromb Haemost. 2008;6(1):176-83.
225. Xu Z, Chen H, Liu D, Huo J. Fibulin-1 is downregulated through promoter hypermethylation in colorectal cancer: a CONSORT study. Medicine (Baltimore). 2015;94(13):e663.
226. Zhu J, Chen R, Mo L, Tang H, Kuang Y, Fei W, et al. Expression of fibulin-1 predicted good prognosis in patients with colorectal cancer. Am J Transl Res. 2015;7(2):339-47.
227. Pio R, Ajona D, Lambris JD. Complement inhibition in cancer therapy. Semin Immunol. 2013;25(1):54-64.

228. Markiewski MM, Lambris JD. Is complement good or bad for cancer patients? A new perspective on an old dilemma. Trends Immunol. 2009;30(6):286-92.
229. Starcevic D, Jelic-Ivanovic Z, Kalimanovska V. Plasma C1 inhibitor in malignant diseases: functional activity versus concentration. Ann Clin Biochem. 1991;28 (Pt 6):595-8.
230. Zeng X, Hood BL, Sun M, Conrads TP, Day RS, Weissfeld JL, et al. Lung cancer serum biomarker discovery using glycoprotein capture and liquid chromatography mass spectrometry. J Proteome Res. 2010;9(12):6440-9.
231. Rutkowski MJ, Sughrue ME, Kane AJ, Mills SA, Parsa AT. Cancer and the complement cascade. Mol Cancer Res. 2010;8(11):1453-65.
232. Chang IW, Lin VC, Wu WJ, Liang PI, Li WM, Yeh BW, et al. Complement Component 1, s Subcomponent Overexpression is an Independent Poor Prognostic Indicator in Patients with Urothelial Carcinomas of the Upper Urinary Tract and Urinary Bladder. J Cancer. 2016;7(11):1396-405.
233. Nordstrom M, Wingren C, Rose C, Bjartell A, Becker C, Lilja H, et al. Identification of plasma protein profiles associated with risk groups of prostate cancer patients. Proteomics Clin Appl. 2014;8(11-12):951-62.
234. Lee MJ, Na K, Jeong SK, Lim JS, Kim SA, Lee MJ, et al. Identification of human complement factor B as a novel biomarker candidate for pancreatic ductal adenocarcinoma. J Proteome Res. 2014;13(11):4878-88.
235. Hu D, Zheng H, Liu H, Li M, Ren W, Liao W, et al. Immunoglobulin expression and its biological significance in cancer cells. Cell Mol Immunol. 2008;5(5):319-24.
236. Wang J, Lin D, Peng H, Huang Y, Huang J, Gu J. Cancer-derived immunoglobulin G promotes tumor cell growth and proliferation through inducing production of reactive oxygen species. Cell Death Dis. 2013;4:e945.
237. Niu N, Zhang J, Huang T, Sun Y, Chen Z, Yi W, et al. IgG expression in human colorectal cancer and its relationship to cancer cell behaviors. PloS one. 2012;7(11):e47362.
238. Kuo WT, Lee TC, Yu LC. Eritoran Suppresses Colon Cancer by Altering a Functional Balance in Toll-like Receptors That Bind Lipopolysaccharide. Cancer Res. 2016;76(16):4684-95.
239. Yesudhas D, Gosu V, Anwar MA, Choi S. Multiple roles of toll-like receptor 4 in colorectal cancer. Front Immunol. 2014;5:334.
240. Li C, Luo X, Lin Y, Tang X, Ling L, Wang L, et al. A Higher Frequency of CD14+ CD169+ Monocytes/Macrophages in Patients with Colorectal Cancer. PloS one. 2015;10(10):e0141817.
241. Baenke F, Peck B, Miess H, Schulze A. Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development. Dis Model Mech. 2013;6(6):1353-63.
242. Hassan MI, Waheed A, Yadav S, Singh TP, Ahmad F. Zinc alpha 2-glycoprotein: a multidisciplinary protein. Mol Cancer Res. 2008;6(6):892-906.
243. Chung L, Moore K, Phillips L, Boyle FM, Marsh DJ, Baxter RC. Novel serum protein biomarker panel revealed by mass spectrometry and its prognostic value in breast cancer. Breast Cancer Res. 2014;16(3):R63.
244. Xue Y, Yu F, Yan D, Cui F, Tang H, Wang X, et al. Zinc-alpha-2-glycoprotein: a candidate biomarker for colon cancer diagnosis in Chinese population. Int J Mol Sci. 2014;16(1):691-703.
245. Bruno R, Olivares R, Berille J, Chaikin P, Vivier N, Hammershaimb L, et al. Alpha-1-acid glycoprotein as an independent predictor for treatment effects and a prognostic factor of survival in patients with non-small cell lung cancer treated with docetaxel. Clinical cancer research : an official journal of the American Association for Cancer Research. 2003;9(3):1077-82.

246. Huang Z, Ung T. Effect of alpha-1-acid glycoprotein binding on pharmacokinetics and pharmacodynamics. Curr Drug Metab. 2013;14(2):226-38.
247. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012;9(2):179-81.
248. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011;19(7):807-12.
249. Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, et al. HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. Proteomics. 2005;5(13):3262-77.
250. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75(17):4646-58.
251. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
252. Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G. Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. Journal of Statistical Software,. 2015;67(6):241-70.