Centro de Investigación en Matemáticas

Unidad Monterrey

Análisis de datos genéticos en población mexicana y su relación con el cáncer colorrectal

T E S I N A

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Cómputo Estadístico

PRESENTA:

Edison Jessie Vázquez Gordillo

ASESOR DE TESIS:
Dr. Rodrigo Macías Páez
CO-ASESOR DE TESIS:
Dr. Augusto Rojas Martinez

Monterrey, Nuevo León, 2018

A Dios por la oportunidad de cumplir este sueño. A mis padres, Hortencio y Maria Elena, por su apoyo incondicional y sus esfuerzos por tener una educacion. A mi prometida Jenny, por su comprensión y sus palabras de aliento que me motivarón a cada instante.

RECONOCIMIENTOS

Primeramente agradezco al Centro de Investigación en Matemáticas Unidad Monterrey por haberme aceptado ser parte de esta gran institución y abierto sus puertas de su seno científico para poder realizar mis estudios de posgrado, así mismo a los docentes que brindaron sus conocimientos y su apoyo para seguir adelante en cada escalón de este desafio.

Agradezco también al CONACyT por el apoyo económico otorgado para la realización de este programa de maestría.

Mi agradecimiento también a mis asesores de tesina, al Dr. Rodrigo Macías por su apoyo y su conocimiento entregado. Al Dr. Augusto Rojas Martinez y la Dra. Rocio Ortiz Lopez por la oportunidad de participar en esta investigación.

Y para finalizar, también agradezco a todos los que fueron mis compañeros de clases durante esto dos años de programa, ya que gracias a la amistad de ustedes los días mas dificiles se pudieron pasar con alegría.

DECLARACIÓN DE AUTENTICIDAD

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesina es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesina es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Edison Jessie Vázquez Gordillo. Monterrey, Nuevo León, 2018

RESUMEN

En la actualidad se han generado bases de datos genómicas para el estudio de la relación de las variantes genéticas humanas y enfermedades; esto implica tratar con bases de datos de alta dimensionalidad provocando problemas a la hora de realizar análisis computacionales y estadísticos que nos permitan entender la población bajo estudio. En este trabajo de titulación se realiza una revisión de literatura, aplicación de métodos computacionales y modelos estadísticos que nos ayuden a obtener resultados óptimos en todos los sentidos, tanto como en el área de la computación y la medicina.

Se obtuvierón gráficas de color que nos muestran el porcentaje de ancestralidad de las poblaciones mexicana (MXL), africana (YRI) y española (IBS), donde se obtuvo que dos de los cuatro SNP's tienen mayor porcentaje de ancestralidad europea (española) lo cual puede dar indicio de una relación entre esta población geográfica y el cáncer colorrectal.

Con los resultados y la metodología llevada a cabo se generó en México un antecedente computacional que permita recrear estos análisis géneticos poblaciones en otro tipos de enfermedades y observar el comportamiento y el impacto poblacional en enfermedades crónicas.

Contenido

CAPÍTULO 1

Introducción y objetivo del trabajo

Esta tesina se centra en el tema de la ancestría genética en población mexicana y su relació con el cáncer colorrectal. Como este tema en concreto no está analizado en la literatura especializada, el principal objetivo de este trabajo es dar una visión de la problemática y presentar el análisis computacional desarrollado.

Por otro lado, la problematica computacional implicada en estos tipos de estudios está relacionada con el gran tamaño del conjunto de datos (volumen) y a su complejidad (alta dimensión), la cuales dificultan su captura, gestión, procesamiento o análisis mediante técnologías y herramientas computacionales.

Por lo tanto, en forma específica, esta tesina tiene como objetivo final definir la relación de ancestría de los genes con la suceptibilidad del desarrollo de cáncer colorrectal a través de métodos computacionales y estadísticos de alto rendimiento. En este trabajo vamos a analizar la ancestría poblacional del grupo a estudiar, obteniendo el porcentaje de genes pertenecientes a una región en especifico y comparar las regiones de genes susceptibles al cáncer colorrectal con las regiones de genes del grupo estudiado, para determinar si hay un vínculo entre la ancestría de estos genes con el desarrollo del cáncer.

Para conseguir estos objetivos se presenta una descripción breve de lo que es el análisis de ancestría genética y sus dependencia o problemática en relación al desarrollo de cáncer. Posteriormente, para definir el porcentaje de ancestría de genes de la población a estudiar, se establecerán los marcadores informativos de ancestría (AIMs) que indican a que región geográfica pertenece cada gen. Por último, enfocaremos el estudio de los métodos para el análisis computacional y estadístico que apoye a lograr el objetivo final.

Así, la finalidad de este tesina es el de lograr generar antecendentes en el estudio de ancestría genética y su relación con el desarrollo de enfermedades cancerígenas y el análisis óptimo de los datos provenientes de estos estudios.

En la actualidad, en el mundo se estima alrededor de 38 millones de muertes anuales, de las cuáles el 63 % de estas difunciones son a causa de enfermedades no transmisibles (ENT) que generalmente son crónicas y de larga duración ya que progresan lentamente. Los cuatro tipos de mayor importancia son las enfermedades cardiovasculares, las enfermedades crónicas, la diabetes y el **cáncer** INEGI (2016).

El cáncer de colon y recto es el cuarto cáncer más frecuente en México y a nivel mundial. De los 38 millones de muerte por las ENT, casi un millón de estas muertes son causadas por este cáncer INSP (2015). Además representa el 2.68 % de todos los tumores malignos Dario et al. (2016).

Por otro lado, las enfermedades cróno-degenarativas como el cáncer, pueden tener relación de desarrollo con los agentes ambientales Weinberg (1988) y con variantes genéticas.

Por tanto, en la conceptualización del problema, existen cuatro puntos a considerar:

(1) Las variantes genéticas están implicadas en la susceptibilidad al cáncer.

(2) Las variantes genéticas en una población dependen de los flujos genéticas que han ocurrido durante las migraciones y los procesos de selección natural a las que se han sometido las poblaciones actuales.

(3) En el caso de México, se dio un proceso historico de mezcla de por lo menos tres poblaciones humanas ancestrales: nativo americano, española y africana.

(4) El proceso de mestizaje se caracteriza por la coexistencia de variantes genómicas ancestrales que puedan tener un afecto en la susceptibilidad al cáncer colorrectal.

Ahora bien, en este contexto, se presenta la primera pregunta a contestar: *¿Hay relación de la ancestría de genes con la suceptibilidad al cáncer colorrectal?*.

En la misma línea del análisis génetico poblacional y su relación con el cáncer colorrectal, se plantea la problemática de la complejidad de la información. Los datos del genoma humano presentan secuencias de variantes géneticas de un solo nucleótido alrededor de *5 millones de organismos*, implicando gran cantidad de reserva en memoria de cualquier computador Beatriz et al. (2006). Este proceso se basa en un esfuerzo de alta magnitud al la hora de procesar o evaluar la evolución, escabilidad y dimensionalidad de las redes transcripcionales de genes Paulino and Antonio (2009).

Los problemas cómunes en los estudios de génetica, con respecto al rendimiento de los métodos computacionales y técnicas estadísticas, citando a Domingo (2006), son:

- **El tiempo de procesamiento**, el cual suele ser elevado al tener que analizar grandes cantidades de datos.
- **Las variables utilizadas**, son de difícil determinación e influyen considerablemente en el resultado final.
- **La medida de asociación**, que es utilizada para el agrupamiento y condiciones, la cual suele ser de gran magnitud.
- **La evaluación** final de los datos y sus interpretación.

Por tanto, la segunda pregunta para esta investigación es *¿Hay capacidad de herramientas computacionales para el procesamiento de los datos complejos y el análisis óptimo de los mismos?*.

En el marco de la investigación de la génetica de población, realizar este estudio dará una panaroma mas amplio en el tratamiento del cáncer colorrectal, de manera similar, sera un antecedente en estudios de ancestralidad en México para diferentes enfemerdades complejas. Además de lograr resultados con aplicación real en este tipo de cáncer.

Por otro lado, la necesidad de comprender grandes y complejos conjuntos de datos, principalmente en el área de la genética, desarrollará una habilidad de extraer conocimiento útil y actuar en consecuencia a este conocimiento extraído, para mejorar los procesos computacionales y estadísticos al momento de evaluar y procesar la información.

El cómputo estadístico es un campo que permitirá, aunado con especialistas en el área a donde se enfoque, a la construcción de nuevos procesos de analisís de datos de alta complejidad en función a los objetivos propuestos.

Así pues, este trabajo de investigación tiene como objetivo principal determinar la asociación entre el cáncer colorrectal y la ancestría de genes por medio de herramientas estadísticas computacionales y evaluaciones médicas.

Por tanto, los objetivos especifícos se dividen en dos áreas:

- **Génomica**:
  - Encontrar el procentaje de ancestría de las regiones geográficas en los individuos del grupo con la enfermedad.
  - Revisar las regiones cromosómicas asociadas con el desarrollo del cáncer colo rectal e identificarlas en los genes de los individuos caso control.
  - Identificar la ancestría de los genes que tuvieron mayor implicación con las regiones cromosómicas asociadas a la enfermedad.

- **Cómputo estadístico:**
  - Comparar softwares computacionales para la lectura y procesamiento de la base de datos génomicos
  - Desarrollar mapas de color para la visualización de la ancestría de lo genes de cada individuo

El contenido de esta investigación se encuentra dividido en los siguientes capítulos:

Trabajos previos. Se introduce algunos trabajos que se han generado en el área de genómica problacional con respecto a la relación de ancestria individual y enfermedades crónicas. Además se describe los procedimientos metodológicos y los resultados que estos han encontrado en sus investigaciones.

Conceptos técnicos y software ADMIXTURE y tratamiento computacional de los datos. Se describen las bases teóricas del software computacional a utilizar. De manera similar, se definen algunos conceptos en el área de genómica poblacional con respecto a la estimación de la ancestría y el tratamiento de los datos.

Elementos probabílisticos y metodología. Se describe los métodos computacionales y estadísticos que tienen como base el software ADMIXTURE que se usan para la estimación para el valor **K** y la ancestría individual de la población estudiada.

Resultados. Se presenta los resultados del análisis, cuáles han sido el porcentaje de ancestría de genes en la población estudiada y además presentar la relación existente o no con el desarrollo del cáncer colorrectal.

Conclusiones y futuros trabajos. Se visualizará el proceso de la discusión de los resultados y se realiza las conclusiones y recomendaciones del proyecto de investigación e investigación futura.

CAPÍTULO 2

Trabajos Previos

A continuación se revisan algunos estudios relacionados al estudio de la implicación de la ancestría de los genes con los riesgos de desarrollar alguna enfermedad.

Veronica et al. (2006) estudia la base génetica de la diabetes tipo 2 (T2D) con el objetivo de identificar los factores de riesgo genéticos y su prevalencia entre los principales grupos continentales. El estudio, es concebido como una investigación de campo tipo explicativa, se concetró principalemente en la ciudad de México, donde se encuentra centralizada la mayor parte de centros de especialización en problemas de obesidad y la diabetes. La recolección de datos de casos y control, se realizó por medio de donantes de sangre sanos que fueron invitados a participar en un estudio para identificar factores de riesgo de diabetes tipo 2. Las muestras de los pacientes con esta enfermedad fueron recolectados entre los años 2000 y 2005 por las Unidades de Investigación Clínica de Bioquímica y Epidemiología del Centro Médico "Siglo XXI". Al final obtuvieron 286 pacientes con T2D (198 mujeres, 88 hombres) y 275 casos control (86 mujeres, 189 hombres). Se genotiparon 69 Marcadores Informativos de Ancestralidad(AIMs) en los casos de control y los casos enfermos. Estos marcadores tienen grandes diferencias de frecuencia entre poblaciones de ascendencia indígena, europea y africana. Para el análisis de las proporciones de ascendencia individual utilizaron el software ADMIXMAP. El análisis de los resultados permitió reconocer las proporciones promedio de contribución de las distintas regiones geográficas definidas en los AIMs en la prevalencia de T2D.

De manera similar, en la consulta de trabajos relacionados con el estudio de la ancestría de los genes y su asociación con algún tipo de cáncer se han observado varias investigaciones tales como el trabajo de Julie (2018) "La génetica del cáncer de seno en la familia puertorriqueña", en el que desarrolla una investigación explicativa en el país de Puerto Rico en 1572 mujeres con cáncer de seno con los objetivos de (1) identificar los factores genéticos involucrados en el riesgo de cáncer en la población puertorriqueña y (2) determinar el rol de la ancestría genética en el desarrollo de los tumores de cáncer de seno agresivo de tipo triple negativo. Para el objetivo (2) utilizaron marcadores informativos genéticos (AIMs) en los cuáles observaron que los puertorriqueños

son mayormente de ascedencia europea y africana, pero menos influencia del genoma taíno que las poblaciones de México. Los resultados muestran que las mujeres hispanas tienen menor riesgo de desarrollar cáncer de seno que las mujeres europeas. Aunque también se observó que las mujeres puertorriqueñas con ascendencia africana tiene más probabilidad de desarrollar este tipo de cáncer.

Marta et al. (2016) realiza un estudio de asociación genómica en una población de ascendencia nativo americana con el riesgo a desarrollar la enfermedad autoinmune de lupus eritemotaso sistémico (LES). Con el objetivo de identificar los loci de riesgo genético para LES se analizarón 3,710 individuos en casos y controles con LES donde para la estimación de ancestría local utilizarón el software PCAdmix con un grupo de ancestría, donde K=3. Los resultados confirmo que los genes principales para el lupus provienen de ascedencia europea y asiática. Lo que al final demuestra que los genes afectados por esta enfermedad están claramente establecidos y asociados a través de etnias.

Justo et al. (2017) estudia las asociaciones de ascendencia mapuche con riesgo de cáncer biliar en la región de Chile. Se enfoca en las posibles asociaciones entre el tipo de ascendencia indígena de esa región y las principales causas de muerte. Con 64 pacientes con cáncer biliar y 170 controles sanos obtenidos de una muestra de 1805 chilenos mezclados y 639,789 muertes, utilizó análisis de componente principales genéticos y estimación de componentes de ancestría usando la función *eigenstrat* del software estadístico R y el programa de Admixture respectivamente. La relación entre la ancestría genética y los datos agregados de mortalidad se analizó con la ayuda del método estadístico de regresión lineal múltiple para estimar las proporciones de ascendencia regional esperadas, además de la regresión múltiple de Poisson para cuantificar la asociación entre las tasas de mortalidad regional y los componentes de ascedencia esperados. Para encontrar la proporción de ancestría en K poblaciones, utilizaron el método computacional Cross-Validation que indicó que el número de poblaciones es igual a cuatro. El análisis de los resultados confirmarón la asociación entre la ascendencia mapache y el cáncer de vesícula biliar.

Nuri et al. (2014) investiga la bacteria Helicobacter pylori como la principal causa de cáncer gástrico. En su estudio menciona que se han utilizado los análisis para la mezcla de poblaciones para describir la historia evolutiva de las bacterias en relación con las migraciones humanas, enfocándose en ampliar este análisis para observar el papel de la mezcla que H. pylori ha jugado en la enfermedad gástrica en los humanos. La recolección de datos se efectúo en dos lugares en el Estado de Nariño, Colombia: desde Tumaco, en la costa del Pacífico y Tuquerres, en la Cordillera de los Andes. Se obtuvier-on 138 hombres y 153 mujeres. Estos presentaban síntomas dispépticos para ser considerados dentro de la muestra. El análisis de ancestria se realizó con el programa STRUCTURE que revelaba que el máximo de poblaciones en la muestra era de K=3, con la distribución de tres grupo estimados de ancestría como europeos, africanos y nativos americanos. También realizó un análisis complementario PCA usando la paquetería

SNPRelate del software R. Los resultados muestran que la bacteria H. pylori es benigno en africanos, mientras que es perjudicial en individuos con ascedencia nativa americana.

S.María et al. (2017) realiza un estudio de asociación de un bloque haplotipos 4-SNP de IL1B con el riesgo de cáncer colorrectal (CRC). La obtención de la muestra se resume en tres grupos de participantes, pólipos adenomatosos (AP), CRC y controles de las regiones andinas y costeras de Colombia. El análisis de ancestría global por individuo fue calculado con el software *Admixture* donde las estimaciones fueron en tres regiones: europeos, africanos y nativos americanos. Para la asociación de la ancestría probaron diferentes análisis de regresión logística multinomial que modelan fenotipos por proporciones de ascedencia global. Los mejores modelos mostrar on que la ascedencia europea se asocia solo con el riesgo de AP, mientras que la ascedencia africana se asocia con ambas enfermedades AP y CRC. Por lo tanto en colombianos con altas proporciones de ascedencia africana en el locus 2q14 albergan más copias de IL1B-CGTC y en consecuencia tienen mayor riesgo de cáncer colorectal.

CAPÍTULO 3

Conceptos técnicos, software ADMIXTURE y tratamiento computacional de los datos

A continuación se presentan las bases teóricas que sustentan la investigación sobre el uso de los métodos computacionales y los métodos estadísticos para el análisis de los datos, y la identificación de la relación de ancestría con el cáncer colorrectal. También una descripción breve del cáncer colorrectal y la estructura genética de México.

3.1 Proceso del cáncer colorrectal

El cáncer en general, se origina por cambios genéticos en las células, estas a su vez son la parte básica que forman los tejidos, los cuales forman los órganos del cuerpo. En este proceso, las células crecen y se dividen para crear nuevas células que se dividen sin control e invaden tejidos, alterando los tejidos formando una masa, que es lo que se conoce como tumor NCI (2008).

Y con respecto al origen del cáncer colorrectal, o también llamado cáncer de colon o de recto dependiendo donde se desarrolle, la American Cancer Society ACS (2014) menciona que la mayoría de estos tipos de cáncer comienzan como un crecimiento llamado *pólipo* en el revistimiento del colon o del recto como se observa en la figura 3.1. Según ASCRS (2018), los pólipos son "crecimientos anormales de tejido que sugen del capa interior o mucosa del intestino grueso (colon) y sobresalen al canal intestinal".

Figura 3.1: Localización del cáncer colorrectal Nuria (2017)

En la mayoria de los casos de cáncer, los factores que estan asociados al desarrollo en estas enfermedades tienen una interacción entre sí. En el cáncer colorrectal, los factores genéticos y ambientales que provocan su desarrollo se favorecen por la interacción de algunos de ellos como el estilo de vida, la dieta y la herencia génetica. En estilos de vida, la falta de ejercicio físico, el sobrepeso y la obesidad, así como el consumo de tabaco y alcohol, son comunes en casos con este tipo de cáncer. De manera similar, existe relación con el consumo de carnes rojas, carne procesada y carne expuesta al fuego directamente, aunque no se ha determinado de que manera los alimentos ricos en fibra, vegetales y leche fungen como protectores ante este tipo de cáncer Nuria (2017). La herencia, representa entre un 20 a un 25 % de los factores de riesgo para esta enfermedad Rodrigo and Riestra (2007). De manera similar, el cáncer colorrectal tiene una prevalencia en individuos mayores a 50 años, siendo esto uno de los mayores riesgos de padecer esta enfermedad.

3.2 Estructura genética de la población mexicana

México es un país ubicado en américa del norte y es el tercer país más grande de América Latina. Al día de hoy, el país esta conformado por 110 grupo étnicos que componen la mayor parte de la población como se muestra en la figura 3.2. La población de meztizos en México es de gran proporción según los datos del INEGI (2016). En general, el concepto de *mestizo* en México es para referirse personas con una aparencia fenotípica intermedia entre los esterotipos europeos o africanos y tipos de indígenas endemicos del país Andrés et al. (2014).

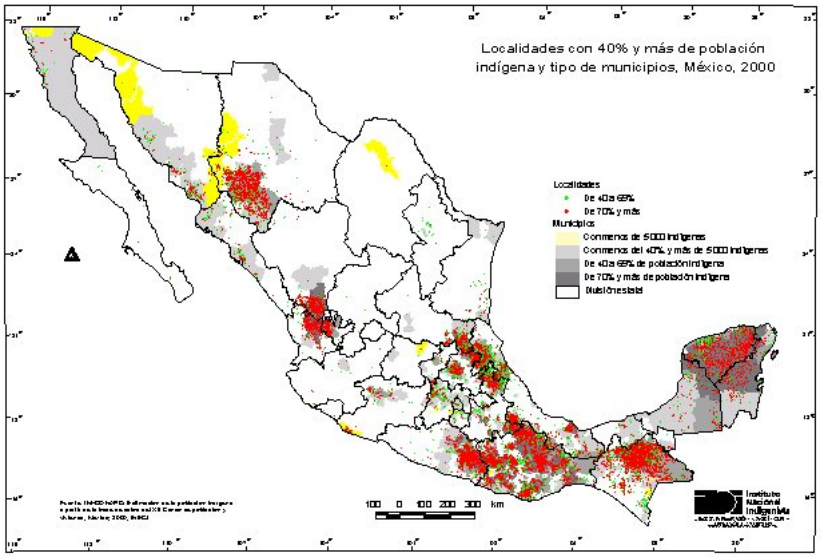


Figura 3.2: Mapa de calor para localidades con 40 % y más de población indígena en México. El color rojo muestra las regiones con presencia indígena mayor a 70 % mientras el verde representa la presencia indígena menor al 40 % Enrique et al. (2006)

La historia de la conquista de México empezó en el año 1559 donde los europeos, sobre todo españoles, arribaron a las costas de Yucatán, empezando el mestizaje en esa zona, añadiendose individuos del Africa traídos en esclavitud a América. Por lo tanto, la mezcla genética entre los americanos nativos (indígenas), europeos (españoles) y africanos que se produjo en México formó una población mestiza con variadas características físicas Carlos (2016); Gabriela and Héctor (2013) como se observa en las siguientes figuras 3.3 y 3.4.

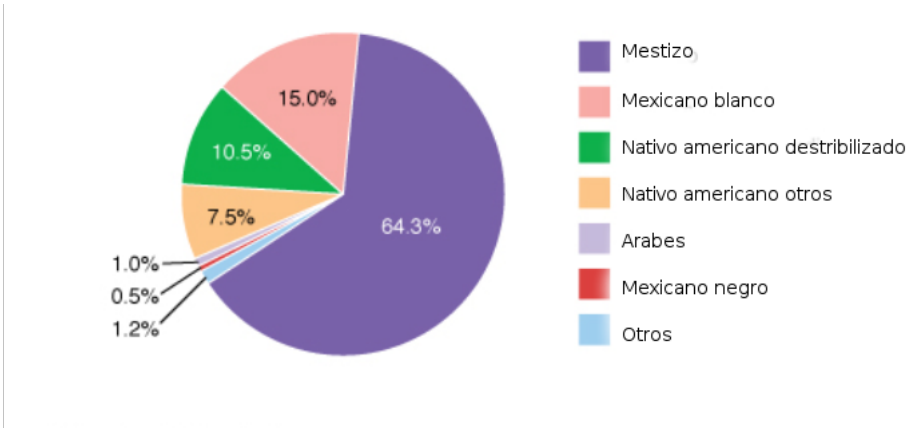


Figura 3.3: Porcentaje de poblaciones por origen genético en México Catherine (2018)

Figura 3.4: Mapa de calor de dos poblaciones con origen genético europeo y americano nativo Gabriela and Héctor (2013)

3.3 Genotipado

La Academia Europea de Pacientes ACP (2015) define el genotipado como "el proceso mediante el cual se determinan las diferencias en los caracteres genéticos o el genotipo de un individuo mediante el análisis de su secuencia de ADN individual. Esto se puede hacer mediante la comparación del genotipo con la secuencia de otro individuo o con una secuencia de referencia". En otras palabras, el genotipado es la técnica de

laboratorio que se utiliza para determinar la información génetica de un organismo, o genotipo, y poder individualizar del resto, así como la susceptibilidad y variantes causales a una enfermedad.

Las bases de datos que capturan el genotipado se encuentran en formato *.ped* y *.map* y de un archivo en formato *.xlsx* que muestra las variantes genéticas por cambio de un solo nucleótido (SNP, en inglés) asociados a la enfermedad. Los archivos .ped describen los individuos y los datos genéticos de la población estudiada. Este archivo puede ser delimitado por *ESPACIOS* o *TAB*, cada línea corresponde a un solo individuo como se puede observar en la siguiente tabla.

FAM 1	IND1	0	0	1	0	A	A	T	T	0	0	...
FAM 2	IND2	0	0	1	0	A	G	T	C	T	A	...
FAM 3	TRIOF	0	0	1	0	A	G	T	C	A	T	...
FAM 4	TRIOM	0	0	2	0	A	G	T	C	A	T	...
FAM 5	TRIOC	TRIOF	TRIOF	1	0	A	A	C	T	A	T	...

Tabla 3.1: Ejemplo estructura datos .ped

Las primera seis columnas son:

1. Family ID [string]
2. Individual ID [string]
3. Father ID [string]
4. Mother ID [string]
5. Sex [integer]
6. Phenotype [float]

Las columnas 7 y 8 codifican los alelos observados en SNP1, las columnas 9 y 10 codifican los alelos observados en SNP2, y así sucesivamente. Los datos faltantes se codifican como "0 0". Este archivo debe tener N líneas y 2L + 6 número de columnas, donde N y L son los números de individuos y SNP contenidos en el conjunto de datos, respectivamente. Es importante resaltar que cada individuo debe tener una identificación única que contenga solo caracteres alfanuméricos.

Por otro lado, el archivo .map describe los SNPs. La estructura de esta información se puede observar en la tabla 3.2.

7	SNP1	0	123
7	SNP3	0	456
7	SNP3	0	789

Tabla 3.2: Ejemplo de estructura de datos .map

De manera similar, este archivo puede estar delimitado por ESPACIOS o TAB. Cada línea corresponde a los SNPs y las cuatro columnas son:

1. Chromosome number [integer]
2. SNP ID [string]
3. SNP genetic position (cM) [float]
4. SNP physical position (bp) [integer]

3.4 Análisis de mezcla de poblaciones

El análisis de mezcla de poblaciones se define, según la International Society of Genetic Genealogy ISGG (2018), como un método para inferir los orígenes geográficos de alguien basado en un análisis de su ascedencia genética. La mezcla (Admixture) sucede cuando las poblaciones comienzan con el mestizaje y la desendencia de estas producen una mezcla de alelos de diferentes poblaciones ancestrales. El análisis de este acontecimiento de ancestralidad tiene una importancia valiosa tanto en génetica de poblaciones como en epidemiología genética Line et al. (2013).

Esta técnica tiene como base la clasificación posible de los individuos en series de grupos étnicos comúnmente identificados tales como europeos, americanos nativos, entre otros. Esto es posible mediante el uso de un SNP informativo de origen ancestral para obtener el porcentaje del genoma de un individuo que tiene un origen ancestral dado Orlando et al. (2016), esto se puede observar en la figura 3.5.

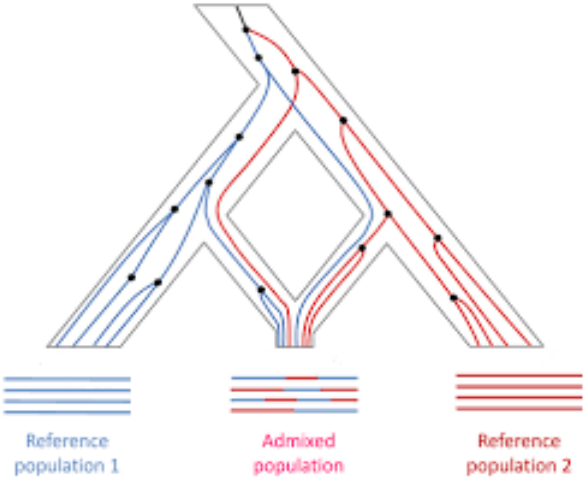


Figura 3.5: "La figura muestra un árbol de población (en gris) y un árbol de genes (en azul y rojo) que rastrea la historia evolutiva de dos poblaciones ancestrales y su correspondiente población mezclada. Los haplotipos específicos de la población de referencia 1 (líneas rojas) podrían fluir y mezclarse con la población 2 y viceversa." Kai et al. (2017)

3.5 Software ADMIXTURE para la estimación de la ancestralidad

Varios autores han usado diferentes métodos para la estimación de ancestría Veronica et al. (2006); Justo et al. (2017); Nuri et al. (2014). Los softwares de mayor uso, especificamente diseñados para realizar *admixture mapping* en la literatura son los programas `STRUCTURE`, `MALDsoft`, `ADMIXMAP`, `ANCESTRYMAP` y `ADMIXTURE`. Este último calcula las estimaciones mucho más rápido usando un algoritmo númerico de optimización. Especificamente ADMIXTURE es un software para la estimación de máxima verosimilitud de ancestros individuales de conjuntos de datos genotipo SNP multilocus. ADMIXTURE utiliza un enfoque de relajación de bloques (**block relaxation**, ver el apartado 4.2) para actualizar alternativamente la frecuencia de alelos y los parámetros de la fracción ascendente. Cada actualización de bloques se maneja resolviendo una gran cantidad de problemas de optimización convexos independientes, que se abordan usando un algoritmo de programación cuadrático secuencial rápido. La convergencia del algoritmo se acelera utilizando un novedoso método de aceleración *cuasi Newton* 4.3. El algoritmo supera a los algoritmos EM y los métodos de muestreo MCMC por un amplio margen Alexander et al. (2009). Una descripción general de los programas para la estimación de ancestría se presenta en la tabla 3.3.

A grandes rasgos, el software ADMIXTURE estima la probabilidad para los genotipos observados basandose en las proporciones de ascedencia y frecuencias de alelos de población, realizándolo simultaneamente. El archivo de entrada debe de representar

Programa	**Global/local**	**Método principal**
STRUCTURE	Global/Local	MCMC:Markov Chain Monte Carlo
frappe	Global	ML: Maximum likelihood
ADMIXTURE	Global	EM: Expectation Maximization
EIGENSTRAT/smartpca	Global	PCA: Principal Component Analysis
ipPCA/EigenDev	Global	PCA: Principal Component Analysis
GEMTools	Global	Spectral graph
PLINK	Global	EM: Expectation Maximization
LAMP	Local/Global	Hierarchical Hidden Markov Model
HAPMIX	Local/Global	Infinite hidden Markov model
ANCESTRYMAP	Local/Global	Bayesian and ML

Tabla 3.3: Descripción de softwares para la estimación de ancestria local o globlal Yushi et al. (2013)

a los genotipos de individuos no relacionados, el cual puede ser una estimación del número de poblaciones `K`.

Ahora bien, para la estimación de la ancestría ADMIXTURE enfoca sus estimaciones en el método de máxima verosimilitud (EM) (ver 4.4), en lugar de los métodos tradicionales para estos tipos de estudio como muestrear la distribución posterior utilizando MCMC. Además, con el método de block relaxation aumenta la velocidad de la estimación haciendo que sea superior en eficiencia computacional respecto a otros programas de alto nivel como STRUCTURE. En líneas generales, ADMIXTURE actualiza el parámetro de frecuencia del alelo y el parámetro de fracción ascendente alternativamente maximizando la expansión de la función de verosimilitud de Taylor de segundo orden. Para realizar este proceso se usa programación cuadrátrica secuencial y se desarrolla de manera iterativa en función de las frecuencias de los alelos y las proporciones de ascendencia asociadas con los valores de los parámetros actuales.Dado que es iterativo y se necesita encontrar un punto óptimo para resolver $x - M(x) = 0$, el método de Newton se puede usar para esta búsqueda. Sin embargo, obtener el diferencial de M(x) es desafiante, por lo que se usa un método cuasi-Newton, lo que permite la acelaración de convergencia y tiene una ventaja sobre los métodos de velocidad sobre convergencia como el método EM. Se ha probado con datos reales y se ha encontrado que ADMIXTURE es mucho más rápido que STRUCTURE pero con una estimación comparable Yushi et al. (2013).

Por otro lado, existen dos denominaciones de ADMIXTURE para estimar la ancestría. La estimación de ascedencia se demonina **Supervisada** cuando se considera los genotipos de individuos con ancestros conocidos, para ello se necesita archivos `.pop`. Cuando no se incluyen individuos con ancestros conocidos entonces se denomina no supervisada Timothy and Justo (2014).

3.6 Estimación de Ancestría por proyección

Cuando se tiene un conjunto de datos nuevo de genes, donde se tiene como objetivo estimar la ancestria, existe la manera de utilizar conjuntos de datos como paneles de referencia, tales como los proyectos **1000Genomes** o **HapMap**. Estos se usan en combinación con la muestra del estudio para estimación de la ancestría utilizando softwares como ADMIXTURE, debido a que estos grandes conjuntos de datos resumen la estructura de la población humana. Es decir, para las muestras del estudio que no incluyen una población nueva, una forma eficiente de estimar la ascedencia individual es **proyectar** las nuevas muestras en la estructura de la población conocida (aprendida) de los paneles de referencia.

La manera en que se realiza esta operación tiene una estructura similar a la operación de proyección utilizada en el análisis de componentes principales, aunque los

detalles matemáticos difieren. Para el tipo de poyección no supervisada se requiere que las dos bases de datos (los paneles de referencia y los datos del estudio) tengan los mismos SNP's, siendo que se aprende del conjunto de datos de referencia y las proporciones de ascendencia se pueden inferir para el conjunto de datos del estudio.

En el aspecto matemático, esto requiere resolver el problema de maximización de verosimilitud de la ecuación 4.2, del capitulo 4, con respecto a `Q` para un `P` fijo. Este tipo de problemas son convexos y se resuelven de manera eficiente con la ayuda de los algoritmos de optimización como Block Relaxation Suyash et al. (2016).

CAPÍTULO 4

Elementos probabilísticos y metodología

En este capítulo, se describe la introducción al desarollo de la tesis, los métodos de computación para el procesamiento de los datos y su análisis. De manera similar, los elementos probabilísticos para el análisis de datos.

4.1 Modelo Probabilístico

Las bases de datos típicas para la estimación de la ancestría consiste en genotipos en un gran número `J` de SNPs de un gran número `I` de individuos no relacionados. Como en cualquier parte del mundo, estos individuos provienen de una población mixta con contribuciones de poblaciones ancestrales postuladas por `K`. La población `k` contribuye con una fracción q_{ik} de un gemona $i's$ individual. Por lo tanto, el alelo 1 en el SNP `j` tiene la frecuencia f_{kj} en la población `k`.

Puede ser que el alelo 1 sea el alelo menor y el alelo 2 sea el alelo principal o viceversa, no importa ya que esto deriva al mismo resultado. Lo que realmente importa es que tanto el la fracción de contribución q_{ik} y la frecuencia f_{kj} son desconocidas. Por lo tanto, es ncesario estimar a q_{ik} para saber la ascendencia en un estudio de asociación , pero también enfocandonos en la estimación de f_{kj}. Esto nos permite estimar el grado de divergencia entra las poblaciones ancestrales estimadas utilizando la estadística F_{ST}.

El modelo estadístico de likelihood que adopta ADMIXTURE viene de STRUCTURE, donde los individuos están formados por la unión aleatoria de gametos, lo que produce las proporciones binomiales

$$\begin{aligned} &Pr(1/1 \text{ para cada i en el SNP j}) = [\textstyle\sum_k q_{ik} f_{kj}]^2 \\ &Pr(1/2 \text{ para cada i en el SNP j}) = 2[\textstyle\sum_k q_{ik} f_{kj}][\sum_k q_{ik}(1 - f_{kj})] \\ &Pr(2/2 \text{ para cada i en el SNP j}) = [\textstyle\sum_k q_{ik}(1 - f_{kj})]^2. \end{aligned} \tag{4.1}$$

Este modelo realiza una suposición adicional de equilibrio de vinculación (linkage equilibrium) entre los marcadores. Además de que los conjuntos grandes o densos de marcadores deben ser podados con el proposito de mitigar el desequilibrio de ligamento (LD) de fondo.

Por otro lado, el registro de los datos se realiza por recuentos. Por lo tanto, g_{ij} representa el número observado de copias de alelo 1 en el marcador `j` de la persona `i`. En consecuencia, g_{ij} puede ser igual a 0,1 o 2 de acuerdo al genotipo 2/2,1/2 o 1/1 de la persona `i` en el marcador `j`. Si consideramos que los individuos son independientes, que para todos los casos así se consideran, la función loglikehood de la muestra entera es

$$L(Q,F) = \sum_i \sum_j g_{ij} ln[\sum_k q_{ik} f_{kj}] + (2 - g_{ij}) ln[\sum_k q_{ik}(1 - f_{kj})]. \tag{4.2}$$

Se observa que los parámetros $Q = q_{ik}$ y $F = f_{kj}$ con dimensiones IXK y KXJ respectivamente, dando un total de $K(I + J)$ parámetros. En un ejemplo real, si consideramos que $I = 1000$, $J = 10,000$ y $K = 3$, se tendrían que estimar alrededor de 33,000 parámetros. Este número hace que el método de Newton no sea factible. El espacio requerido para la matriz Hesiana es demasiado grande, y su matriz inversa de esta es computacionalmente costosa.

4.2 Algoritmo de Relajación de Bloques

El algoritmo de relajación de bloques es diferente a los métodos de descenso de coordenadas, los cuales tienen la gran ventaja de que conducen a problemas de optimización unidimensional, siendo muchos más sencillos que los multidimensionales. En la mayoría de los casos reales tenemos problemas multidimensionales como en el área de la génetica. A continuación se dará una definición breve del algoritmo de relajación de bloques(BR) Leeuw (1994).

El método de relajación de bloques se define matematicamente como métodos de punto fijo. De manera general se da una descripción de los métodos de punto fijo. Se llama un punto fijo a un punto que satisface la siguiente ecuación.

$$x = g(x) \tag{4.3}$$

El teorema del punto fijo define a `D` como un conjunto, y busca condiciones en `D` y `g` que garantizen la existencia de `x`; condiciones que garantizan unicidad (no obligatoriamente). Para ello se puede utilizar el método de iteracción de punto fijo que se muestra en la tabla 4.1.

Algoritmo Fixed Point Iteration
1. Dada una ecuación f(x) = 0
2. Convertimos la ecuación f(x)=0 a la forma x = g(x)
3. Damos un valor inicial aleatorio para x_0
4. Do $$x_{i+1} = g(x_i)$$
5. while(no se cumpla ninguno de los dos criterios de convergencia `C1` o `C2`)

Tabla 4.1: Descripción algoritmo Fixed Point Iteration

- `C1`: Se corrigió apriori el número total de iteracciones N
- `C2`: Al probar la condición $|x_{i+1} - g(x_i)|$ (donde **i** es el número de iteracciones) con un limite de tolerancia ϵ, donde se fija el apriori.

Entendiendo el concepto del método punto fijo, se consideró la siguiente condición general para el algoritmo de relajación de bloques. Se minimizo la función **f** de valor real en el conjunto de productos $X = X_1 \otimes X_2 \otimes ... \otimes X_p$, donde $X_s \subseteq \mathsf{R}^{n_s}$.

Para minimizar esta función se utiliza el siguiente método iterativo (Tabla 4.2, que tiene su base en el algoritmo iterativo anterior 4.1.

Comenzar:	Empezar con $x^{(0)} \in X$
Paso k.1:	$x_1^{k+1} \in \underset{x_1 \in X_1}{\operatorname{argmin}} f(x_1, x_2^k, ..., x_p^k)$.
Paso k.2:	$x_2^{k+1} \in \underset{x_2 \in X_2}{\operatorname{argmin}} f(x_1^{k+1}, x_2, x_3^k ..., x_p^k)$.
...	...
Paso k.p:	$x_p^{k+1} \in \underset{x_p \in X_p}{\operatorname{argmin}} f(x_1^{k+1}, ..., x_{p-1}^{k+1}, x_p)$.
Motor: k $\leftarrow$ k + 1 e ir hacia k.1	

Tabla 4.2: Método iterativo

En este método iterativo se puede observar que existen mínimos en la subetapas, pero no necesitan ser únicos (aunque se espera que esta condición exista). El argumento de mínimo son mapas punto a punto, aunque en muchos casos se asignan en singletons (conjunto con exactamente un elemento). En los problemas reales se harán cálculos con

selección del argmin.

4.3 Acelaración de Convergencia

Como es bien conocido, los algoritmos EM no son en su defecto de altas tasas de convergencia. De manera similar, en el esquema de relajación de bloques, aunque es más rápido que los algoritmos EM, aún tiene poca potencia de convergencia, por tanto es necesario utilizar un acelador de convergencia. A continuación se describe un método génerico que se usó junto con el software ADMIXTURE .

Se supone que un algoritmo está definido por un mapa de iteración $x^{n+1} = M(x^n)$. Dado que el punto óptimo es un punto fijo del mapa de iteración, uno puede intentar encontrar el punto óptimo aplicando el método de Newton a la ecuación $x - M(x) = 0$. Debido a que la diferencial $dM(x)$ muestra resistencia al computar, entonces los métodos cuasi-Newton buscan aproximarlo mediante condiciones secantes que involucran repeticiones previas. Para mantener la complejidad computacional bajo control, se limita el número de condiciones de la secante durante la aceleración, así evitando un sobreajuste de operaciones. Además este método tiene las ventajas de evitar el almacenamiento y la inversión de matrices grandes y preservar las restricciones lineales de igualdad. Para mantener la complejidad computacional bajo control, se limitó el número de condiciones de la secante durante la aceleración. La propiedad de ascenso del algoritmo EM y la relajación del bloque son útiles para monitorear la aceleración. Cualquier paso acelerado que lleve cuesta abajo es rechazado a favor de un paso ordinario. Los pasos acelerados no dependen necesariamente de las restricciones, por lo que las actualizaciones de los parámetros están cayendo fuera de sus regiones factibles Zhou et al. (2011).

4.4 Algoritmo EM

A través del programa de ADMIXTURE se usó el algoritmo EM de FRAPPE que básicamente actualiza los parámetros a través de lo siguiente:

$$f_{kj}^{n+1} = \frac{\sum_i g_{ij} a_{ijk}^n}{\sum_i g_{ij} a_{ijk}^n + \sum_i (2 - g_{ij}) b_{ijk}^n}, \tag{4.4}$$

$$q_{ik}^{n+1} = \frac{1}{2J} \sum_j [g_{ij} a_{ijk}^n + (2 - g_{ij}) b_{ijk}^n], \tag{4.5}$$

por simplicidad y conveniencia lo definimos de la siguiente forma,

$$a^n_{ijk} = \frac{q^n_{ik} f^n_{kj}}{\sum_m q^n_{im} f^n_{mj}}, b^n_{ijk} = \frac{q^n_{ik}(1-f^n_{kj})}{\sum_m q^n_{im}(1-f^n_{mj})}$$

como es conocido, los algoritmos EM son de lenta convergencia y el que utiliza el programa FRAPPE no es la excepción, por eso se usa la aceleración de convergencia. Aunque es necesario realizar y conocer el estado o diagnosticar la convergencia, una manera simple es declarar la convergencia una vez que los loglikelihoods susecivas cumplan lo siguiente

$$L(Q^{n+1}, F^{n+1}) - L(Q^n, F^n) < \epsilon, \tag{4.6}$$

donde el software ADMIXTURE usa en principio un valor epsilon igual a 10^{-4}, mientras que para FRAPPE el valor de ϵ es 1. Esto es diferente dado que ADMIXTURE ha tenido mejores estimaciones con un epsilon pequeño o muy menor a 1. Por tanto, se determinó que el valor de ϵ es 10^{-4}.

4.5 Tratamiento computacional de los datos

Como se mencionó en el apartado 3.3, las bases de datos en este estudio están en formato *.ped* y *.map*. En un primer acercamiento se visualizó la dimensión de los datos. El archivo .map tiene `1,006,658` registros (SNPs), y cuatro columnas. Mientras que el archivo .ped consta de `1712` (881 CASOS CON CCR Y 831 CASOS CONTROLES SANOS, DE LOS CUALES LA PROPORCIÓN DE GÉNERO ES 1012 HOMBRES Y 700 MUJERES) registros, que pasarón el control de calidad, con `6,893,628,847` columnas. El peso de los archivos son de `25 Mb` y `6.9 Gb` respectivamente.

Los archivos contienen la información de los genotipos de los individuos genotipados en el proyecto CHIBCHA. Los datos fueron expuestos a sucesivos controles de calidad en base a diversos criterios técnicos y en base a parámetros poblacionales de la población mexicana.

El procesamiento de los datos y su análisis se realizó mediante el apoyo del servidor perteneciente al **CIMAT unidad Monterrey**. La capacidad del servidor es de `32 Gb` en memoria RAM y `8` núcleos. Se utilizó la configuración de cómputo en threads (4 threads) para realizar un menor tiempo de análisis.

4.6 Validación Cruzada y la estimación del Parámetro K

El software ADMIXTURE permite elegir el número de poblaciones ancestrales, **K**. Este número es realmente importante y en muchos casos no sabemos de cuántas poblaciones ancestrales han descendido nuestras muestras.

Por tanto, para estimar el número de poblaciones (**K**) que representará a la muestra es necesario tomar en consideración las siguientes suposiciones:

- Suponemos que hay **K** poblaciones ancestrales $A_1, ..., A_K$ que se han estado mezclando por **g** generaciones.
- Las poblaciones ancestrales son desconocidas y pueden implicar frecuencias de alelos diferentes en cada repetición

Tomando en cuenta las anteriores suposiciones y entendiendo que existen K poblaciones en nuestra muestra, se consideró utilizar el método de validación cruzada que nos permite evaluar la capacidad predictiva de los posibles modelos para ayudar a determinar el número adecuado de componentes que se deben conservar en el modelo, en nuestro caso los componentes se reduce a obtener el valor K. Además, el método de validación cruzada es la mejor opción si no se sabe cuál es el número óptimo de componentes (número de poblaciones ancestrales).

Esta estimación para identificar el "mejor" valor para K (número de poblaciones) se puede realizar de diferentes formas, por ejemplo, uno de los programas con mayor referencia en estos tipos de estudio *STRUCTURE* (http://pritch.bsd.uchicago.edu/structure.html) usa un modelo de evidencia para K definido como:

$$Pr(G|K) = \int f(G|Q, P, K)\pi(Q, P|K)dQdP$$

aproximando esta integral por el método Monte Carlo combinado con una distribución apriori no informativa a través del Teorema de Bayes para obtener las probabilidades posteriores Pr(k—G).

Por otro lado, ADMIXTURE utiliza el procedimiento de *Cross-validation* para identificar el valor de **K**. Este procedimiento divide los genotipos observados en v=5 (por defecto) folds de aproximadamente el mismo tamaño. El procedimiento enmascara (es decir, convierte a "MISSING") todos los genotipos, para cada fold a su vez. Es decir, para cada fold, el conjunto enmascarado G resultante es usado para calcular las estimaciones $\tilde{\theta} = (\tilde{Q},\tilde{P})$. Cada genotipo enmascarado g_{ij} se predice por

$$\hat{\mu}_{ij} = E[g_{ij}|\tilde{Q}, \tilde{P}] = 2\Sigma_k \tilde{q}_{ij}\tilde{P}_{kj},$$

y el error de predicción es estimado por el promedio de los cuadrados de la desvianza residual para el modelo binomial, a través de todas las entradas enmascaradas sobre todos los folds Yushi et al. (2013); Alexander et al. (2009).

$$d(n_{ij}, \tilde{\mu}_{ij}) = n_{ij}log(n_{ij}/\mu_{ij}) + (2 - n_{ij})log[(2 - n_{ij})/(2 - \tilde{\mu}_{ij})] \tag{4.7}$$

Por tanto, dado la situación etnica de México y el conocimiento de la historia del mismo, se decidió realizar cinco ejecuciones y obtener el error de validación para estas posibles poblaciones ancestrales en nuestra muestra. Además, dado la dimensión de los datos y la complejidad de la misma no se quiso realizar otras inferencias de número de

poblaciones que no tuvieran mayor relevancia para esta primera búsqueda.

Cada ejecución fue inicializada con una random seed de valor `43`, por lo que nos ayuda a que cada inferencia para cada rango no se tome distintos bloques de poblaciones. El valor delta para la convergencia fue de `0.0005`, este valor ya viene predeterminado por el mismo programa aunque da la oportunidad de poder cambiarlo. En nuestro caso no fue necesario ya que se ha probado que este valor es muy bueno para inferir el número de poblaciones. El número de iteraciones para la convergencia y el tiempo de corrida se muestra en la tabla 4.3

# de poblaciones	# de iteracciones	Tiempo de ejecución (min)
2	33	311
3	59	645
4	73	922
5	94	1322
	TOTAL	3200 (53 hrs.)

Tabla 4.3: Ejecución para la búsqueda del mejor K

Al obtener el valor del número de poblaciones (**K**=3), se procedió a realizar la estimación de ancestría con las bases de datos `.ped` y `.map` aunado con las bases de datos del proyecto `1000Genomes` en formato PLINK.

4.7 Fst de Wright

La distribución empírica F_{ST} es uno de los tres estadísticos F, también conocidos como índices de fijación, que se usó para describir el nivel esperado de heterocigocidad en las tres poblaciones estudiadas. El concepto de heterocigocidad se define al heredar dos formas diferentes de un gen en particular, una de cada progenitor NHGRI (2014).

De manera similar, Wright (1965) define al estadístico F_{ST} como la correlación de alelos (variantes de un gen) extraídas al azar de la misma población en relación con la población total, donde la población total puede verse como la combinación de dos muestras de poblaciones.

Aunque el estadístico F también puede ser definido como una medida de la correlación entre genes muestrados a diferentes niveles de una población subvidida jerárquicamente, los autores Gaurav et al. (2013) han tomado en consideración que la definición

más apegada en génetica de poblaciones es la que menciona Weir and Hill (2002) "como la correlación entre los alelos extraídos aleatoriamente de una sola población en relación con la población ancestral común más reciente",

$$E[p_i^s|p_{anc}^s] = p_{anc}^s \tag{4.8}$$

$$Var(p_i^s|p_{anc}^s) = F_{ST}^i p_{anc}^s(1 - p_{anc}^s) \tag{4.9}$$

donde p_i^s es la frecuencia alélica del alelo derivado de la población i, en el SNP s, mientras que p_{anc}^s es la frecuencia alélica del alelo derivado de la población ancestral en el SNP s, y F_{ST}^i es la población especifica F_{ST} para la población i. Por ejemplo, para un par de poblaciones, la distribución F_{ST} es,

$$F_{ST} = \frac{F_{ST}^1 + F_{ST}^2}{2} \tag{4.10}$$

4.8 Estimación de la relación con el cáncer colorrectal

En el caso de la estimación de ancestría se definió en primera instancia el tipo de análisis a usar. En nuestro caso, al no conocer las poblaciones apriori, una forma eficiente de estimar la ascendencia individual es proyectar nuestras muestras a la problación aprendida(frecuencias alélicas) aprendidas de los paneles de referencia. Este tipo de analisis de le denomina `"Projection Analysis"`, el cual toma como referencia una base de datos de proyectos de genomas ya establecidos como 1000Genomes, HapMap en combinación con la muestra del estudio para estimar la ancestralidad usando el método no supervisado de ADMIXTURE.

Anteriormente, el equipo de trabajo de Uruguay, en el mismo contexto, obtuvo 75 SNPs que mostrarón asociación con la predisposición de producir cáncer colorrectal y que pasaron una prueba multesting, de los cuales 12 SNPs (4.4) tuvierón un valor significante en la prueba Bonferroni y se tomarón en cuenta como los más adecuados para el estudio, ya que muestran mayor asociación con la enfermedad.

SNP	Probabilidad de asociación	Bonferroni	FDR
rs7311395	8.164E-11	0.00009224	0.00009224
rs118184226	5.696E-10	0.0006435	0.0003218
rs55885037	1.844E-09	0.002083	0.0005488
rs2598121	1.943E-09	0.002195	0.0005488
rs7197593	2.573E-09	0.002907	0.0005708
rs115600951	0.000000003	0.003425	0.0005708
rs74382455	6.845E-09	0.007734	0.001105
rs111445080	1.035E-08	0.01169	0.001343
Affx-18048474	1.07E-08	0.01208	0.001343
rs117982396	0.000000018	0.02032	0.002032
rs74455361	2.381E-08	0.0269	0.002445
Affx-17135896	2.66E-08	0.03006	0.002505

Tabla 4.4: Representación SNPs con mayor asoaciación

Conociendo los SNPs, gracias al apoyo del equipo de trabajo de Uruguay, se procedió a realizar la extracción de estos SNPs en el proyecto 1000Genomes. Pero antes, se decidió realizar un rango de ± 100 por posición genetica, por ejemplo, para el SNP `rs11798239` con posición génetica `10:30303271`, donde el 10 representa el cromosoma en el que esta y el 30303271 es locación, se tomarón los SNPs en el rango de 30303271 ± 100; esto con el propósito de conocer los SNPs mas cercanos al SNP relacionado.

Considerando lo anterior, se realizó la búsqueda de rangos en la base de datos de 1000Genomes (`http://www.internationalgenome.org/data`). Conociendo el número de cromosomas de los SNPs fue mas fácil su búsqueda, además de contar con su posición génetica. De los 12 SNPs, se observarón solo 3 SNPs en su completa referencia. Es decir, solo podiamos proyectar cuatro SNPs (rs74382455, rs2598121, rs7311395, rs7197593) con la base de datos 1000Genomes. Las poblaciones que se relacionaron fueron las siguientes 4.5.

Población	Descripción Población
YRI	Yoruba in Ibadan, Nigeria
IBS	Iberian Population in Spain
MXL	Mexican Ancestry from Los Angeles USA

Tabla 4.5: Tipo de poblaciones en nuestro estudio

De manera similar, se obtuvo las medidas de correlaciones (F_{ST}) de los genes muestrados de las poblaciones de las cinco ejecuciones.

El preprocesado de los archivos de 1000Genomes constaba en convertirlos a formato PLINK (.ped y .map), eliminar los SNPs repetidos, y recodificarlos. Esto se realizó con la ayuda de la paquetería `vcftools,` http://vcftools.sourceforge.net/.

Al contar con estas bases de datos o archivos de referencia, se realizó un join para juntar todos los archivos en uno solo. Después se compararon el número de SNPs con la base de datos CHIBCHA .Ped para eliminar los SNP's sobrantes.Al final de los `1,006,658` SNPs registrados, se preservaron **1203 SNP's**. Esto nos permitió evaluar el trabajo de cómputo de mejor manera y evitar un sobreajuste en nuestras estimaciones.

Al correr el primer análisis de estimación de ancestría se observó que era necesario realizar uno donde solo se tomarán en cuenta los **casos** con cáncer, ya que es la población de mayor relevancia en nuestro estudio. Además de dividir la población en hombres y mujeres.

El preprocesado de datos, la descarga de las bases de datos de referencia (1000Genomes) y la corrida del programa ADMIXTURE se realizó por medio de la terminal de Linux.

CAPÍTULO 5

Resultados

La figura 5.1 muestra el error cuadrático de estimación de grupos en la población CHIBCHA para toda la muestra, sin haber eliminado ningún SNP. Este proceso de Cross-Validation muestra que en nuestra base de datos existe tres grupos, esto puede justificarse con la historia de conquista en México.

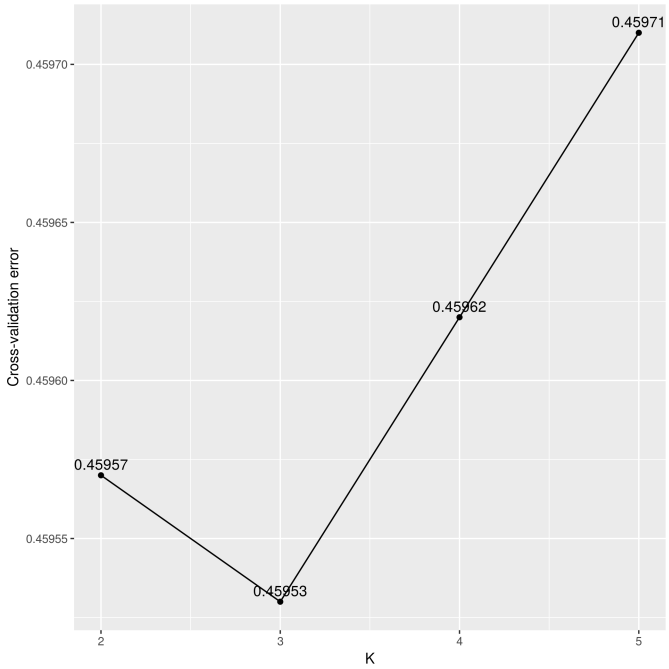


Figura 5.1: Estimación de K, número de poblaciones posibles en la base de datos CHIBCHA.

	pob0
pob0	
pob1	0.109

Tabla 5.1: Fst divergencia entre las poblaciones estimadas para K=2

	pob0	pob1
pob1	0.017	
pob2	0.121	0.082

Tabla 5.2: Fst divergencia entre las poblaciones estimadas para K=3

	pob0	pob1	pob2
pob1	0.035		
pob2	0.089	0.061	
pob3	0.119	0.078	0.015

Tabla 5.3: Fst divergencia entre las poblaciones estimadas para K=4

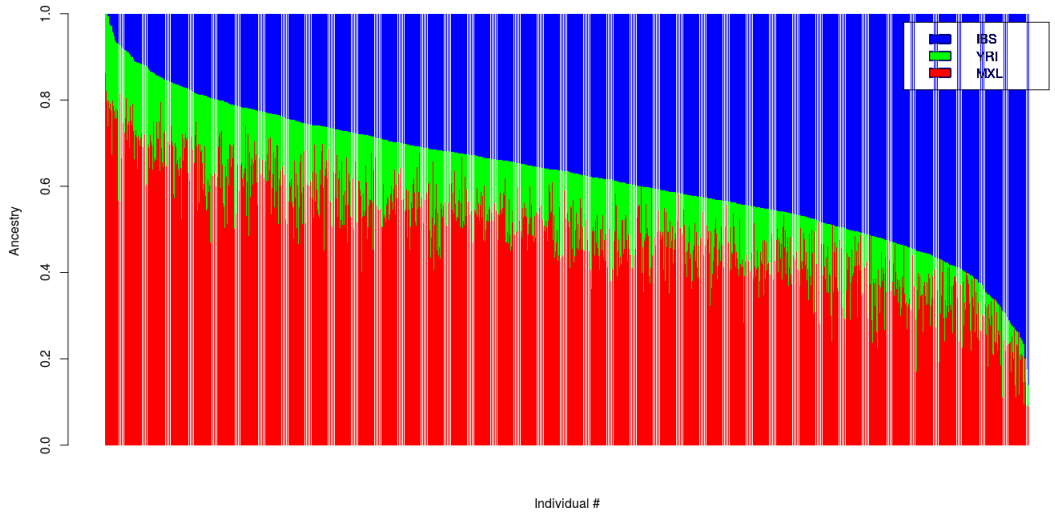
	pob0	pob1	pob2	pob3
pob1	0.067			
pob2	0.050	0.055		
pob3	0.017	0.082	0.058	
pob4	0.92	0.041	0.057	0.121

Tabla 5.4: Fst divergencia entre las poblaciones estimadas para K=5

Las tablas de divergencias nos muestran la separación entre las poblaciones estimadas. En este caso en particular se enfoca en la tabla 5.2 donde se observa que la población 2 y la población 0 tienen el valor de correlación más alta, lo cual puede indicar que en estas poblaciones existen alelos iguales proviniente de una población

ancestral.

Por otro lado, en la figura 5 podemos observar la derivación de las tres poblaciones utilizadas para este estudio (española, méxicana y africana). De manera similar podemos visualizar que la mayoría de los individuos tienen poca frecuencia de la población africana, esto comprueba lo que se viene observando a través de la historia de la conquista en México.



Análisis individual de Ancestría total (casos y controles). Cada individuo es representado por una barra vertical en la gráfica.

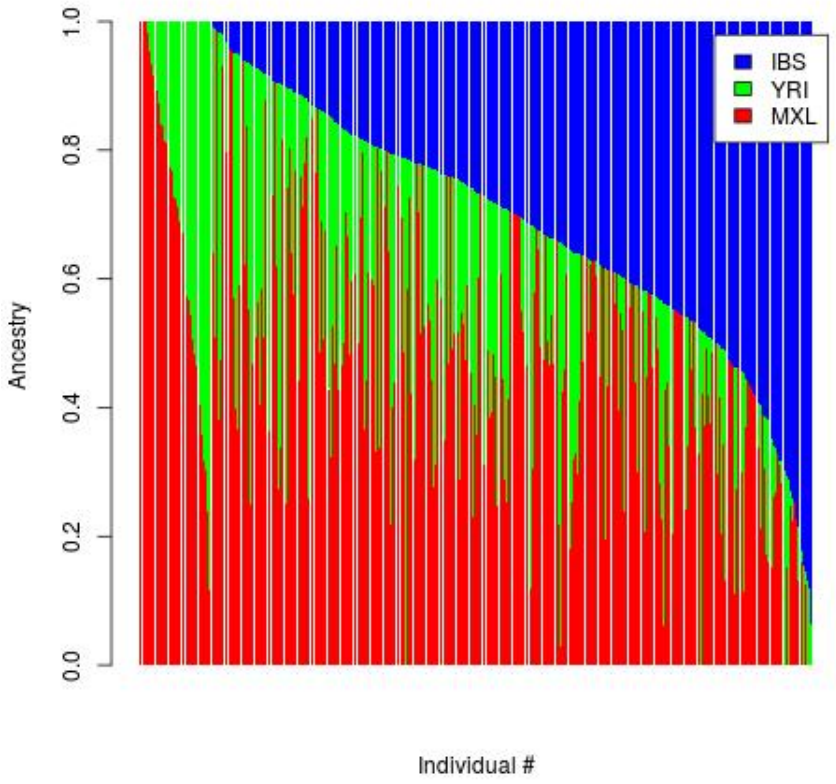


Figura 5.3: Análisis individual de ancestría para casos. Cada individuo es representado por una barra vertical en la gráfica.

La figura 5.3 tiene individuos con 100 % ancestralidad mexicana y casos de individuos sin ningún porcentaje de genes con ancestralidad mexicana. Además de mostrar una diferencia, con respecto al crecimiento del color verde (población africana), con la figura 5. Esto puede deberse a que los casos provienen de entidades como Veracruz o Guerrero dado porcentaje de ancestrilidad altos de la población ancestral africana. De manera similar, dado que los individuos no muestran un distribución heterogénea de colores, puede inferirse que la relación con genes asociados al cáncer colorrectal estará distribuido equitativamente en las tres poblaciones.

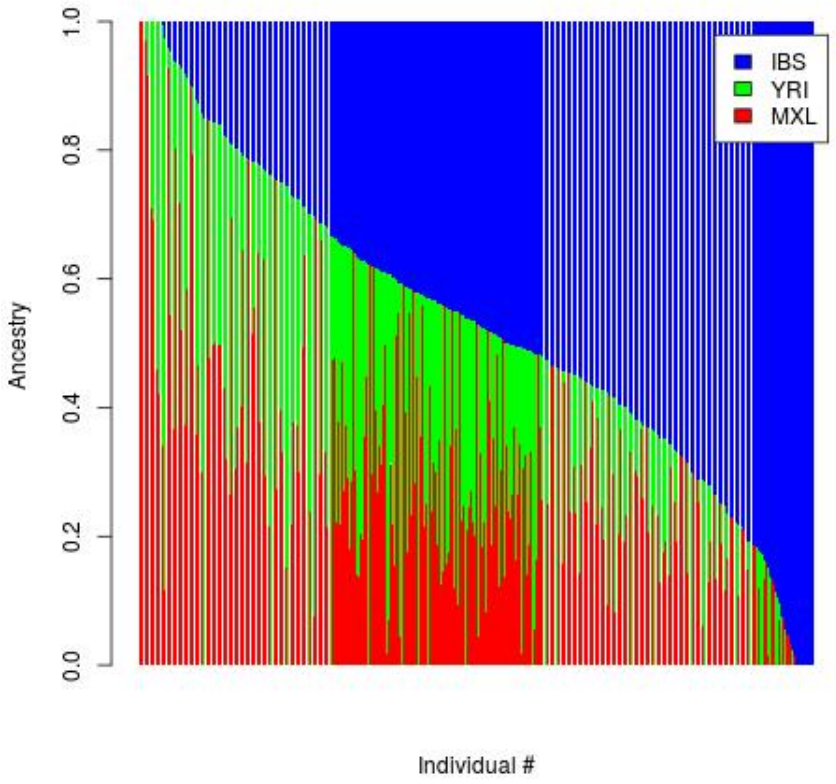


Figura 5.4: Análisis individual de ancestría para casos en género masculino. Cada individuo es representado por una barra vertical en la gráfica.

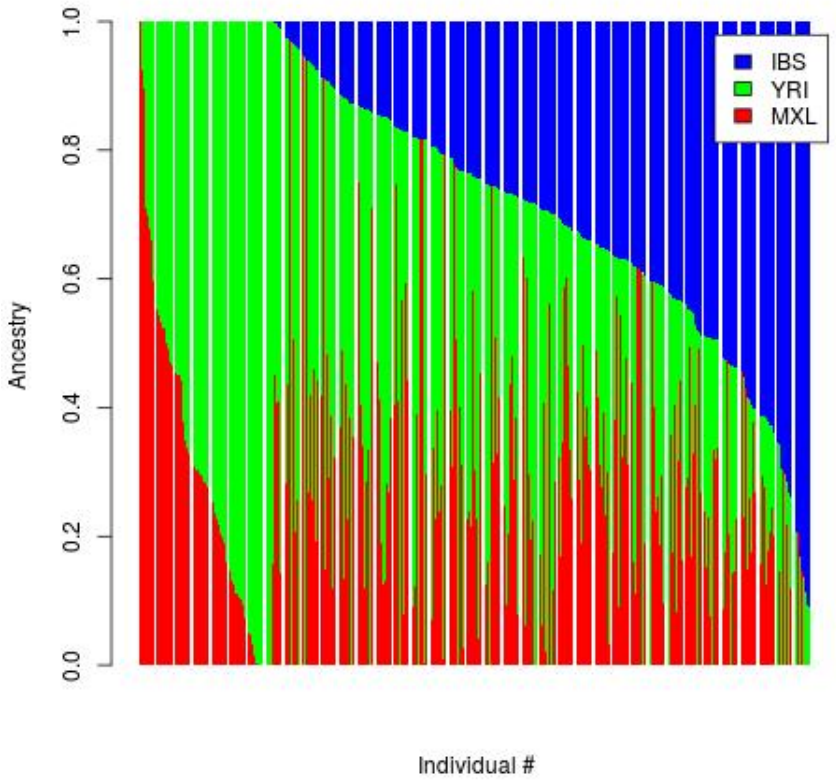


Figura 5.5: Análisis individual de ancestría para casos en género femenino. Cada individuo es representado por una barra vertical en la gráfica.

Con respecto a la posición génetica de los genes, y observando a cada indivudo en la gráfica 5.3, se obtuvo lo siguiente,

SNP	Locación	Población génetica (%)
rs74382455	1:84373659	Africana = 0.90
rs2598121	7:37936286	MXL = 0.96
rs7311395	12:43574178	Europea = 0.86
rs7197593	16:55405713	Europea = 0.97

De cuatro SNPs evaluados, dos de ellos tienen mayor porcentaje de población génetica en españoles. Esto no implica que la población ancestral europea sea la de mayor

correlación, si no que en estos cuatros SNP's tuvieron mayor relación. Puede ser que en los otros SNP's que no se mapearon pueda existir una ventaja de la misma población europea u otra de las dos.

Con respecto a las figuras 5.4 y 5.5 se puede observar que en el cuadro del género femenino existen mujeres con un porcentaje del 100 en genes con ascedencia Africana, de manera similar, se puede observar que en el cuadro de hombres, hay individuos con porcentajes de 100 en ascedencia europea, esto siendo muy marcado en los distintos cuadros.

CAPÍTULO 6

Conclusiones y futuros trabajos

En la actualidad, el cómputo y las herramientas estadísticas forman gran parte para entender el comportamiento del mundo y de la vida. Los resultados que hemos encontrado, en colaboración con el Laboratorio Nacional de Medicina en Sistemas y el equipo de trabajo de Uruguay, han descrito un antecedente en la búsqueda de respuestas antes enfermedades como el cáncer que son un gran problema en el sector de salud nacional.

Los SNPs mencionados están divididos en las tres poblaciones, pero muestran mayor actividad en la población europea, básicamente en la española. Esto puede implicar que los genes asociados al cáncer colorrectal tiene un relación con la ascedencia europea, dado el mestizaje en México, aunque como se mencionó en el apartado de resultados, falta observar el comportamiento de los otros SNPs que no se mapearon y dar un resultado más completo. Por otro lado, siendo que en los cuadros 5.5 y 5.4 se ve una diferencia en porcentaje de genes para cada población, es notable mencionar que hay mujeres con ascedencia americana nativa al 100 % y de manera similar con ascedencia africana pero no existe ningún individuo con ascedencia europea; esto puede implicar que tal vez exista una relación mayor de la enfermedad con estas poblaciones, pero este no es el caso.

Estos análisis tienen como objetivo dar una explicación más contudente, aunque es necesario recalcar la complejidad computacional que esta presenta, ya que las bases de datos son de alta dimensionalidad y su análisis generan problemas complejos. Por otro lado, es interesante notar que de los doce SNPs con mayor asociación al cáncer solo se pudo observar cuatro, esto puede resolverse en la forma de generar mayores estudios en la parte génetica poblacional aquí en México y tener antecedentes que nos ayuden a entender el por que de las enfermedades.

6.1 Futuros trabajos

Aunque se ha generado un gran avance en la estimación de la ancestría en un conjunto de datos de genes con poblaciones desconocida, es necesario recalcar que este tipo de análisis no indica con precisión la cercania de la relación de la ancestria con el cáncer ya que se tiene un rango en la posición génetica provocando no conocer puntualmente la posición del gen. Por lo tanto, en un trabajo futuro se recomienda comenzar a trabajar con análisis de haploides el cuál puede generar mayor referencia en los asuntos de precisión para relacionar las enfermedades con los genes.

Lista de Figuras

Lista de Tablas

Bibliografía

INEGI. Estadisticas a proposito del...dia mundial contra el cancer(4 de febrero). *Cubos dinamicos*, 2016. 2, 9

INSP. Cancer de colon y recto. *https:// www.insp.mx/infografias/cancer-colon-recto.html*, 2015. 2

B. Dario, M. Martin, C. Miguel, et al. Epidemiology of colorectal cancer in patients under 50 years old in the hospital juárez of méxico. *Endoscopia*, 24, 2016. 2

R. Weinberg. Finding the anti-oncogene. *Scientific American*, 259, 1988. 2

P. Beatriz, R. Domingo, and D. Norberto. Analisis de datos de expresion genetica. *Jornadas de Automatica*, 2006. 2

M. Paulino and F. Antonio. *Genetica y Genomica en Acuicultura*. Fundacion Observartorio Español de Acuicultura, 2 edition, 2009. 2

R. Domingo. *Análisis de datos de Expresión Genética mediante técnicas de Biclustering*. PhD thesis, Universidad de Sevilla, 2006. 2

M. Veronica, V. Adan, C. Emily, et al. Admixture in mexico city: implications for admixture mapping of 2 diabetes genetic risk factors. *Hum Genet*, 120, 2006. 5, 14

D. Julie. La genética del cáncer de seno en la familia puertorriqueña. *Revista Puertorriqueña de Medicina y Salud Pública*, 2018. 5

A. Marta, Z. Julie, M. Julio, et al. Genome-wide association study in an amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of european admixture. *Arthritis Rheumatology*, 68, 2016. 6

L. Justo, B. Felix, G. Rosa, et al. Subtypes of native american ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in chile. *PLoS Genet*, 13, 2017. 6, 14

K. Nuri, P. Alvaro, G. Barbara, et al. Human and helicobacter pylori coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences*, 111(4), 2014. 6, 14

S.María, H. Gustavo, U. Adriana, et al. Il1b-cgtc haplotype is associated with colorectal cancer in admixed individuals with increased african ancestry. *Scientific Reports*, 7, 2017. 7

NCI. Lo que usted necesita saber sobre el cáncer de colon y recto. *National Cancer Institute*, 2008. 8

ACS. Colorectal cancer facts figures 2014-2016. *American Cancer Society*, 2014. 8

ASCRS. Pólipos del colon y el recto. https://www.fascrs.org/patients/disease-condition/polipos-del-colon-y-el-recto, 2018. 8

R. Nuria. Comer más proteínas y menos verduras protege (a veces) del cáncer de colon. http://www.abc.es/sociedad/, 2017. 9

L. Rodrigo and S. Riestra. Diet and colon cancer. *Revista Española de Enfermedades Digestivas*, 99, 2007. 9

M. Andrés, R. Christopher, F. Juan Carlos, et al. The genetics of mexico recapitulates native american substructure and affects biomedical traits. *National Institutes of Health*, 344, 2014. 9

S. Enrique, G. Verónica, M. Ismael, et al. Regiones indígenas de méxico. *Programa de las Naciones Unidas para el Desarrollo*, 1, 2006. 10

S. Carlos. Mestizaje y características físicas de la población mexicana. *Arqueología Mexicana*, 65, 2016. 10

M. Gabriela and R. Héctor. El impacto del mestizaje en méxico. *Investigación y Ciencia*, 445, 2013. 10, 11

M. Catherine. Los 50 grupos Étnicos de méxico principales. https://www.fascrs.org/patients/disease-condition/polipos-del-colon-y-el-recto, 2018. 11

ACP. Genotipado. https://www.eupati.eu/es/glossary/genotipado/, 2015. 11

ISGG. Admixture analyses. https://isogg.org/wiki/Admixture$_a$*nalyses*, 2018. 13

S. Line, S. Thorfinn, and A. Anders. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3), 2013. 13

M. Orlando, P. Afshin, I. Tamara, et al. Genetic african ancestry and markers of mineral metabolism in ckd. *CJASN*, 11, 2016. 13

Y. Kai, Z. Ying, N. Xumin, et al. Models, methods and tools for ancestry inference and admixture analysis. *Quantitative Biology*, 5(3), 2017. 14

D. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 2009. 14, 23

L. Yushi, N. Toru, L. Shuguang, et al. Softwares and methods for estimating genetic ancestry in human populations. *Hum Genomics*, 7(1), 2013. 15, 16, 23

A. Timothy and L. Justo. Local and global ancestry inference, and applications to genetic association analysis for admixed populations. *Genetic Epidemiol*, 38, 2014. 16

S. Suyash, D. Carlos, L. Kenneth, et al. Efficient analysis of large datasets and sex bias with admixture. *SOFTWARE*, 2016. 17

De Leeuw. Block relaxation algorithms in statistic. *Information Systems and Data Analysis*, 1994. 19

H. Zhou, D. Alexander, and L. Kenneth. A quasi-newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing*, 21, 2011. 21

NHGRI. Herocigoto. *www.genome.gov/glossarys*, 2014. 24

S. Wright. The genetical stuctrure of populations. *Evolution*, 15, 1965. 24

B. Gaurav, P. Nick, et al. Estimating and interpreting fst: the impact of rare variants. *Genome Research*, 2013. 24

BS. Weir and WG. Hill. Estimating f-statistics. *Annu Rev Genet*, 36, 2002. 25