



UNIVERSIDAD NACIONAL DE COLOMBIA

**Ivan Dario Castillo Abril**

Universidad Nacional de Colombia  
Facultad de Ciencias, Departamento de Estadística  
Bogotá, Colombia  
2011



# Metodología para la construcción de redes de SNPs relacionados con enfermedades complejas

Ivan Dario Castillo Abril

Tesis presentada como requisito parcial para optar al título de:  
**Magister en Ciencias Estadística**

Directora:  
Ph.D. Liliana Lopez Kleine

Línea de Investigación:  
Bioestadística  
Grupo de Investigación:  
Métodos en Bioestadística

Universidad Nacional de Colombia  
Facultad de Ciencias, Departamento de Estadística  
Bogotá, Colombia  
2011



A mis padres

Merceditas y Carlos, que me enseñaron que con perseverancia, trabajo duro y mucho amor en el corazón, se pueden alcanzar las metas en la vida.

Y a Monica

Que me enseñó el valor del amor verdadero y a permanecer firme por lo que se quiere.



## Agradecimientos

Quiero expresar mi agradecimiento a la profesora Liliana Lopez, por su infinita paciencia y sus palabras de apoyo. Sin su respaldo y asesorías con sus observaciones y sus soluciones pragmáticas no hubiera sido posible esta tesis.



## Resumen

El estudio de asociación entre enfermedades complejas y genotipos con base en marcadores genéticos o SNPs (Polimorfismos de Nucleótidos Simples) por sus siglas en ingles, ha tomado importancia en los últimos años. En algunos casos la cantidad de SNPs a analizar es enorme, lo cual dificulta la detección de asociación entre ellos y el fenotipo de interes, por ejemplo, el desarrollo de una enfermedad. Para estos estudios es esencial usar clases de SNPs informativos, que representen al resto de la población. Se requieren metodologías que permitan caracterizar los SNPs de la mejor forma posible para disminuir la cantidad de falsas asociaciones que se puedan detectar. En este trabajo se presenta una metodología utilizando representaciones del espacio de SNPs y metodologías de análisis multivariado para definir clases de SNPs informativos, para luego describir asociaciones entre enfermedades utilizando representantes de cada clase formado una red entre estos.

**Palabras clave:** Polimorfismos de Nucleótidos Simples, GWAS, Genotipos, Fenotipos, Haplotipos, Clases de SNPs, K-means, Red de SNPs.

## Abstract

The study of associations between complex diseases and genetic markers or SNPs (Single Nucleotide Polymorphisms) has become important in recent years. In some cases the SNPs quantity to be analyzed is huge, which impedes the detection of association between some of them and the phenotype (disease). For these studies it is essential to use informative SNPs classes representing accurately to rest of the population. Methodologies are needed to characterize SNPs as best as possible to decrease the number of false associations that can be detected. This work presents a methodology using representations in the SNPs space and the multivariate analysis to define informative SNPs classes, and describe associations between diseases forming a network between them.

**Keywords:** Single Nucleotide Polymorphisms, GWAS, Genotypes, Phenotypes, Haplotypes, SNP's Classes, K-means, SNP's network.

# Contenido

<b>Agradecimientos</b>	<b>vii</b>
<b>Resumen</b>	<b>ix</b>
<b>1. Introducción</b>	<b>2</b>
<b>2. Marco Teórico</b>	<b>5</b>
2.1. Bases de datos de SNPs . . . . .	7
2.2. Estudios de asociación del genoma completo (GWAS) . . . . .	8
2.3. Antecedentes de los estudios de asociación GWAS . . . . .	9
<b>3. Materiales y Métodos</b>	<b>11</b>
3.1. Estructura de datos . . . . .	12
3.2. Tratamiento de datos y métodos . . . . .	13
3.2.1. Métricas sobre $\mathbb{Z}_3^n$ . . . . .	14
3.2.2. Algoritmos de Implementación . . . . .	18
3.2.3. Comparación de poblaciones . . . . .	20
<b>4. Metodologías de Representación y Conglomeración de SNPs</b>	<b>22</b>
4.1. Escalamiento Multidimensional en el espacio de SNPs . . . . .	22
4.2. EMD-NM para la muestra de fenotipos . . . . .	25
4.2.1. EMD-NM para el fenotipo Diabetes tipo II . . . . .	26
4.3. Análisis de conglomerados por agrupamiento K-Means . . . . .	28
4.3.1. K-Means para el fenotipo Diabetes tipo II . . . . .	30
<b>5. Construcción de Redes de SNPs entre Fenotipos</b>	<b>34</b>
5.1. Construcción de redes de SNPs para un mismo fenotipo . . . . .	34
5.2. Relación física de SNPs a lo largo del cromosoma . . . . .	36
<b>6. Discusión General de los Resultados Obtenidos</b>	<b>39</b>
<b>7. Conclusiones y recomendaciones</b>	<b>41</b>
7.1. Conclusiones . . . . .	41
7.2. Recomendaciones . . . . .	41

<b>8. EMD-NM para los fenotipos Enfermedad de Gallstone e Hipertensión</b>	<b>43</b>
<b>9. K-Means para los fenotipos Enfermedad de Gallstone e Hipertensión</b>	<b>49</b>
<b>Bibliografía</b>	<b>54</b>

# 1 Introducción

Los estudios de asociación del genoma completo han tomado gran importancia en los últimos años, estos permitirán identificar sectores del genoma que describan la evolución y posible tratamiento de enfermedades. Actualmente, estos estudios se realizan utilizando marcadores genéticos muy particulares conocidos como Polimorfismos de Nucleótido Simple (SNPs), estos son posiciones en el genoma que varían de individuo a individuo. Los avances que se puedan lograr a partir de las caracterizaciones genéticas de enfermedades serán de utilidad para muchos sectores enfocados en la salud y seguridad social. La industria farmacéutica será uno de los sectores más beneficiados con los resultados positivos de estos estudios [13].

El desarrollo de una enfermedad es el resultado de complejas interacciones entre múltiples factores ambientales y variantes alélicas de muchos genes. En los últimos 30 años, los estudios genéticos de enfermedades multifactoriales humanos han identificado aproximadamente 50 genes y sus variantes alélicas [21] que están asociados al desenlace de enfermedades complejas.

En enero de 2008, el Instituto Nacional de Salud de Estados Unidos (NIH: National Institute of Health) implementó una política para el intercambio de información obtenida de los llamados estudios de asociación del genoma completo (GWAS). El propósito de esta política es fomentar la ciencia en beneficio del público a través de la creación de un repositorio centralizado de datos de Polimorfismos de Nucleótido Simple (SNPs), el NIH-GWAS. El intercambio de información genómica permite a la ciencia médica entender mejor las necesidades de salud de la población y facilitar el desarrollo de nuevas tecnologías y enfoques para la prevención, diagnóstico y tratamiento de enfermedades[7].

En la búsqueda de genes asociados a enfermedades humanas, un paso crucial es detectar las asociaciones entre variantes genéticas y fenotipos de enfermedades, las variantes genéticas están codificadas en haplotipos para cada individuo. Los haplotipos son secuencias de alelos en un cromosoma que ofrecen un marco natural para la realización de análisis conjuntos de múltiples marcadores. Los haplotipos de varios SNPs, son las variables explicativas para los cuales se busca asociación con la variable respuesta (desenlace de una enfermedad), en un grupo de personas enfermas y sus controles, en la mayoría de los GWAS. Así, a través de regresión logística o pruebas  $\chi^2$  se identifican SNPs asociados a la variable binaria desenlace de la enfermedad [1].

---

El proyecto internacional HapMap tiene como objetivo desarrollar un mapa de haplotipos del genoma humano, que describa patrones comunes en la variación de la secuencia del ADN. Con este proyecto se espera que sirva como recurso clave para encontrar genes que afectan la salud, las enfermedades, respuesta a fármacos y factores ambientales [9], ya que al identificar, asociaciones de algunos SNPs, se están indirectamente, identificando genes implicados en la enfermedad en los cuales están ubicados estos marcadores.

Unos 10 millones de SNPs existen en las poblaciones humanas, donde el alelo más raro SNP tiene una frecuencia de al menos 1%. Una región del cromosoma puede contener muchos SNPs, pero sólo unos pocos SNPs “etiqueta” que pueden proporcionar la mayoría de la información sobre el patrón de variación genética en la región [9].

En los últimos años los GWAS se han centrado en ofrecer soluciones al problema de la alta dimensión de los datos, es por esto, que se han desarrollado distintas metodologías estadísticas paramétricas y no paramétricas, para encontrar un conjunto de marcadores genéticos que describa a la población total. En muchos de estos estudios se expresa el espacio de SNPs como un espacio vectorial, obviando muchas veces las condiciones que deben cumplir los elementos del espacio para que tenga la estructura de espacio vectorial. Aunque los resultados son satisfactorios al encontrar SNPs informativos o etiquetas, muchas veces no se sabe si tales etiquetas pertenecen al espacio de SNPs o si es posible asignarles un lugar físico en el cromosoma. Además, en pocos estudios se tiene en cuenta el contexto multivariado, el cual podría aportar información importante sobre la relación de los SNPs con enfermedades y de ellos entre sí.

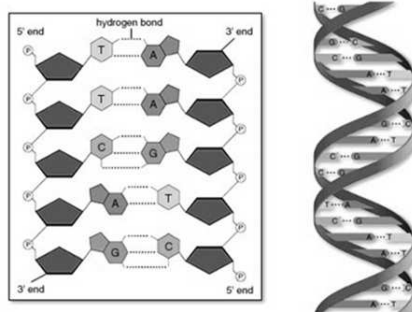
En este trabajo se presenta una nueva metodología basada en resultados obtenidos en la teoría de la información, definiendo una relación de cercanía entre los elementos del espacio de SNPs por medio de distancias de edición. Esto permite luego representar los SNPs provenientes de uno o más fenotipos en un espacio euclidiano usando metodologías de escalamiento multidimensional y ya en este último paso, se pueden encontrar los SNPs etiqueta conociendo su representación espacial identificando conglomerados de elementos del espacio. Una representación de este tipo permitirá análisis adicionales a la simple asociación entre SNPs y el fenotipo. Primero, se reduce el espacio de SNPs de manera objetiva y multivariada con base en su correlación con los fenotipos y no, con base en pruebas múltiples que podrían estar sujetas a la detección de falsos positivos, garantizando que el poder de las pruebas aumente. Segundo, se detectan relaciones entre SNPs que permiten comprender mejor el fenómeno biológico. Tercero, se podrá observar la relación de SNPs con varios fenotipos al tiempo, lo que permitirá detectar cuáles tienen implicaciones en más de una o solamente en una enfermedad.

Este trabajo está organizado en cinco capítulos como sigue. El primer capítulo presenta el

contexto teórico de los estudios de asociación del genoma completo. En el segundo capítulo se presentan los materiales y métodos utilizados para definir la primera fase de la metodología, esto es, asignar una estructura de espacio métrico al conjunto de SNPs para luego definir una relación de cercanía o distancias entre elementos del espacio. En el tercer capítulo se presentan las metodologías de escalamiento multidimensional para encontrar una representación de los SNPs en un espacio euclidiano y de esta forma identificar conglomerados y SNPs etiquetas. En el cuarto capítulo se construyen las redes de SNPs entre conglomerados de un fenotipo particular y luego utilizando los representantes de clase se construye una red global de SNPs entre enfermedades. En este capítulo también se muestra la relación de los conglomerados desde el punto físico dentro del cromosoma. Finalmente en la última parte se concluye con los resultados obtenidos.

## 2 Marco Teórico

El ADN (Acido Desoxirribonucleico) es una molécula compuesta de segmentos llamados nucleótidos. Cada nucleótido consiste de un fosfato, un azúcar y una base nitrogenada. Estas bases nitrogenadas se dividen en dos grupos, las bases purínicas (adenina [A], guanina [G]) y las bases pirimidínicas (citosina [C], timina [T]). La molécula de ADN se presenta como una doble cadena de nucleótidos, en la que las dos hebras están unidas entre si por enlaces conocidos como puentes de hidrógeno, la adenina [A] se enlaza con la timina [T] mediante la formación de dos puentes de hidrógeno, y la guanina [G] se enlaza con la citosina [C] mediante la formación de tres puentes de hidrógeno (fig 2-1)<sup>1</sup>, este patrón se conoce como *apareamiento de bases complementarias* [19].



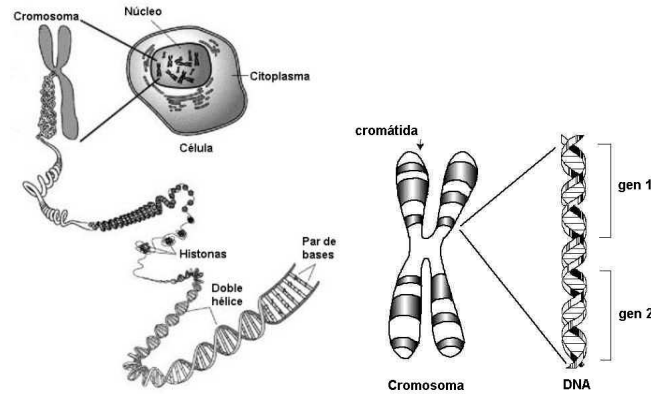
**Figura 2-1:** Estructura en doble hélice del ADN

En humanos el ADN de doble hélice con bases apareadas está presente en el interior del núcleo de las células conformando los cromosomas, los cuales son compendios de información genética. Cada uno de ellos se encuentran agrupados por secciones conocidas como genes (fig 2-2)<sup>2</sup>. Éstos determinan las características hereditarias de la célula u organismo

Cuando se habla de características físicas como por ejemplo, color de ojos, color de cabellos, tono de piel, etc. Estas se conocen como *fenotipo*. A cada fenotipo le corresponde un *genotipo*, ya que las características físicas dependen de los genes. Por tanto, el genotipo o conjunto de genes determina el fenotipo o conjunto de características físicas.

<sup>1</sup>Imagen tomada de Wikipedia

<sup>2</sup>Imagen tomada de Wikipedia



**Figura 2-2:** Cadena de ADN formando un Cromosoma y Genes dentro del Cromosoma

Cada cromosoma que está en el núcleo de la célula es resultado de la combinación de un cromosoma materno y un cromosoma paterno, perteneciendo cada hebra de ADN de uno de los padres. El *locus* de un cromosoma es el lugar físico dentro del cromosoma donde se encuentra una información genética dada o gen. Para un gen específico por ejemplo color de ojos, es posible que tenga la misma información o diferente información. Las variantes que existen de un gen en un mismo locus entre un cromosoma materno y paterno se conocen como *alelos*. Los alelos o variantes alélicas son las posibles opciones para un gen específico, en este caso, color de ojos.

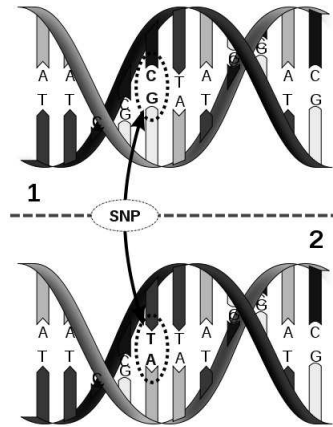
Supongamos que se tiene un alelo color de ojos azules (A) y otro alelo color de ojos marrones (M), estas son las variantes alélicas. En nuestro ejemplo, el alelo (A) se hereda de la madre y el alelo (M) se hereda del padre.

Sea  $C_H$  el cromosoma que es el resultado de la combinación del cromosoma materno  $C_M$  y el cromosoma paterno  $C_P$ . Tomemos el alelo  $g_H$  presente en  $C_H$ . Si  $g_H$  es el mismo para  $C_M$  y  $C_P$  en el mismo locus, diremos que el alelo es *homocigoto*, en nuestro ejemplo  $g_H \in \{AA, MM\}$ , en caso contrario diremos que el alelo es *heterocigoto* esto es,  $g_H \in \{AM, MA\}$ .

Un *polimorfismo genético* es una variación en un lugar determinado del ADN entre individuos de una población. Cuando el polimorfismo consiste en la sustitución de una simple base nitrogenada (adenina [A], timina [T], citosina [C] o guanina [G]) da origen a un *polimorfismo de nucleótido simple* (SNP) (fig 2-3)<sup>3</sup>. La mayor parte de la variación genética entre diferentes individuos de una población se puede caracterizar por SNPs donde las mutaciones en las posiciones de los nucleótidos individuales se produjeron durante la historia humana y se transmite a través de la herencia [8]. Una de estas variaciones debe darse al menos en un 1% de la población para ser considerada como un SNP. Si no se llega al 1% no se considera

<sup>3</sup>Imagen tomada de Wikipedia

SNP sino una mutación puntual. Para una mejor comprensión del genoma humano se ha dispuesto el almacenamiento de millones de SNPs en bases de datos públicas. Por ejemplo el *International HapMap Project*.



**Figura 2-3:** Polimorfismo de Nucleótido Simple (SNP)

## 2.1. Bases de datos de SNPs

El proyecto internacional HapMap es un catálogo de variantes genéticas comunes que ocurren en los seres humanos. Describe cuando se producen estas variantes en el ADN, cómo se distribuyen entre las personas dentro de las poblaciones y entre poblaciones de diferentes partes del mundo. El proyecto está diseñado para proporcionar la información que otros investigadores pueden utilizar para vincular las variantes genéticas con el riesgo de enfermedades específicas, lo que dará lugar a nuevos métodos de prevención, diagnóstico y tratamiento de enfermedades [9].

Mediante la identificación de *haplotipos*, el HapMap proporciona una herramienta que puede ser utilizada en estudios de asociación. En estos estudios, se comparan los haplotipos de un grupo de personas con presencia de enfermedad (casos) y los de otro grupo de personas sin la enfermedad (controles). Si un haplotipo particular, ocurre con más frecuencia en los individuos afectados en comparación con los controles, un gen que influye en la enfermedad puede estar ubicado dentro o cerca de ese haplotipo [9].

Un haplotipo, es una combinación de alelos de diferentes locus en el mismo cromosoma. En otro sentido, un haplotipo es un conjunto de SNPs en un cromosoma particular que están estadísticamente asociados. El número de SNPs identificados es demasiado grande y actualmente se estima que su número llega cerca de los 10 millones en el genoma humano [4].

## 2.2. Estudios de asociación del genoma completo (GWAS)

Un GWAS (*Genome Wide Association Study*), se define como un estudio en el cual la densidad de marcadores genéticos y el grado de desequilibrio entre estos es suficiente para capturar una gran proporción de variaciones comunes en el genoma dentro de una población de estudio.

Los GWAS se realizan utilizando las más recientes herramientas tecnológicas de investigación disponibles, para analizar de forma rápida y rentable las diferencias genéticas entre las personas con enfermedades específicas. El propósito de estos estudios es explorar la conexión entre genes específicos, conociendo la información del genotipo, y su respectivo fenotipo. Esto facilita la identificación de factores de riesgo que describen el desarrollo o progresión de la enfermedad. Recientemente, se han puesto en marcha nuevas iniciativas para realizar GWAS, con la expectativa de que los resultados acelerarán el desarrollo de mejores herramientas de diagnóstico y el diseño de nuevos tratamientos, seguros y eficaces [7].

Análisis cuantitativo y modelización matemática han sugerido que los estudios de asociación del genoma completo resultan ser más eficaces cuando se realizan con *polimorfismos de nucleótido simple* (SNPs). Este tipo de estudios con SNPs representan una ventaja, ya que son fáciles de describir, son muy abundantes y estables [17].

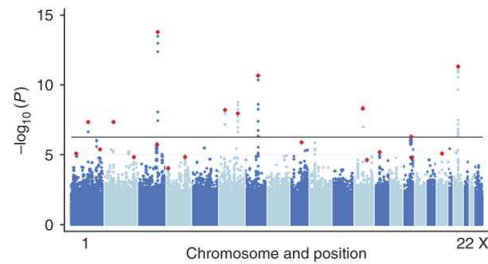
Los estudios casos-control es la vía más utilizada para encontrar asociaciones entre enfermedades debido a la facilidad de recolección y caracterización de los datos. Una metodología clásica para encontrar asociaciones entre marcadores de las muestras casos-controles es la regresión logística. Al considerar una población de marcadores independientes, se define una muestra aleatoria de casos (enfermos) y controles (sanos) de la población, se asigna 1 al fenotipo “Afectado” y 0 “No afectado”, el método busca la estimación de parámetros  $\{\beta_0, \beta_1, \dots, \beta_n\}$  los cuales están asociados a un factor ambiental o variable explicativa, como por ejemplo, edad, sexo, tensión arterial, etc. La variable respuesta en este caso presencia o no de la enfermedad es binaria. La probabilidad de afección es  $p$ , por tanto tenemos

$$p = \frac{1}{1 + \exp(z)} \text{ con } z = \beta_0 + \sum_{i=1}^N \beta_i x_i + \epsilon$$

Donde  $N$  es el número de factores ambientales o variables explicativas para la presencia o no de la enfermedad.

Otra forma de encontrar asociaciones entre marcadores es definiendo una estadística que depende del conteo de alelos presentes en los SNPs de la muestra de casos y controles. Un supuesto sobre la estadística, es que su distribución es  $\chi^2$  luego se le asigna el  $p$ -valor

correspondiente y esta dispersión se grafica, donde el eje horizontal es la distancia física a lo largo del genoma y el eje vertical es el  $-\log(p)$ , los puntos que sobresalen del promedio, son los SNPs relevantes para el estudio de asociación (fig 2-4)[22].



**Figura 2-4:** Representación de asociaciones de SNPs entre posiciones en el cromosoma.

## 2.3. Antecedentes de los estudios de asociación GWAS

El objetivo de los estudios de asociación es identificar patrones de polimorfismos que varían sistemáticamente entre los individuos con diferentes cuadros de enfermedad y poder representar los efectos de los factores de riesgo por medio de la identificación de alelos y así encontrar la forma de describir la enfermedad y su evolución [1].

Debido al gran volumen de información, es necesario seleccionar SNPs informativos (SNPs etiquetas) que representen la distribución original del conjunto global en el genoma. Estos SNPs usualmente son escogidos de bloques de haplotipos conocidos como, haplotipos etiquetas. Este método busca minimizar el número de SNPs para ser utilizados en estudios de asociación. Este subconjunto SNPs tiende maximizar las variaciones genéticas presentes en la población total. La eficiencia de este método depende del análisis estadístico usado. En la práctica el proceso de etiquetado es efectivo para capturar variantes comunes [1].

Halperin et al en [8] definen un algoritmo para encontrar el SNP representante de un haplotipo. A partir del conjunto de SNPs en el haplotipo, una de las hipótesis que usan para definir el algoritmo es que la correlación entre SNPs es directamente proporcional a su posición física en el cromosoma, esto es, la correlación entre SNPs es alta si son muy cercanos entre sí. Zelikovsky et al en [13] define un espacio vectorial de SNPs expresando cada marcador como elemento de  $\mathbb{Z}_3$  y aplicando técnicas de regresión lineal múltiple encuentra los parámetros sobre el espacio vectorial asociado a un haplotipo particular y de esta forma construir de forma cerrada un SNP etiqueta que describa al haplotipo. Jun Zhang et al en [11] utilizan la misma idea de definir un espacio vectorial de SNPs en  $\mathbb{Z}_3$  y la selección de SNPs etiquetas se realiza con una metodología de aprendizaje no supervisado conocido como (SVM) *Support Vector Machines*. Estos autores son los que han hecho mayores aportes al

problema de la búsqueda de SNPs etiqueta, explorando varias técnicas estadísticas desde los modelos paramétricos clásicos de regresión lineal, hasta, modelos modernos no paramétricos de aprendizaje de máquina.

Otros autores que utilizan metodologías no paramétricas para etiquetar SNPs son Chuang et al en [4] donde sobre el espacio de SNPs aplican la metodología (KNN) *K-Nearest Neighbor* que es una versión más simple que la metodología SVM.

### 3 Materiales y Métodos

Siguiendo lineamientos de un estudio de asociación se disponen dos poblaciones de SNPs casos - controles asociados al fenotipo que determina el desenlace de una enfermedad. Se construyen dos conjuntos de SNPs para cada fenotipo, cada conjunto corresponde a la población sana y afectada respectivamente. Sobre las filas de estos conjuntos o haplotipos se disponen los marcadores SNPs y en las columnas se disponen las personas que hacen parte de las muestras casos - controles respectivamente.

A cada una de las personas pertenecientes a las dos poblaciones (sanas y enfermas) se extrae su información genética describiendo la sucesión de alelos presentes en cada uno de los cromosomas. Como los marcadores SNPs tienen asignada una única posición dentro del cromosoma, lo que varía es el alelo correspondiente a cada persona de las dos poblaciones. Un ejemplo de esta disposición se muestra en la tabla **3-1**.

**Tabla 3-1:** Estructura de haplotipos.

SNP	Posición	$P_1$	$P_2$	$P_3$	$\dots$	$P_n$
rs10399749	45162	CC	CC	CT	$\dots$	TC
rs4030303	72434	GG	AG	GG	$\dots$	GG
rs4030300	72515	CC	TC	TC	$\dots$	CC
rs940550	78032	TT	TC	TT	$\dots$	TC
rs13328714	81468	CC	CC	TT	$\dots$	TC
rs11490937	222077	AA	AA	GG	$\dots$	GA
rs6683466	524446	CC	CC	CC	$\dots$	CC
rs12025928	536560	GG	GA	AG	$\dots$	GG

Al no disponer de datos reales de poblaciones enfermas y sanas asociadas a una enfermedad o fenotipo particular, se simulan con datos obtenidos del HapMap, asumiendo que dos poblaciones étnicas son respectivamente los casos y controles. Luego, para cada SNP se realiza un conteo de alelos, homocigotos y heterocigotos asignándoles un valor en  $\mathbb{Z}_3 = 0, 1, 2$  donde 0

es el alelo homocigoto de menor frecuencia, 1 el alelo homocigoto de mayor frecuencia y 2 el alelo heterocigoto. Esta representación de los SNPs determinan cadenas de elementos en  $\mathbb{Z}_3$ , utilizando distancias de edición es posible definir una estructura de espacio métrico a estos elementos.

La posibilidad de medir los elementos del espacio de SNPs, permite la definición de una medida de correlación, de tal forma que se pueden encontrar asociaciones entre las poblaciones étnicas, que es nuestro caso particular y para darle un contexto se asociación entre enfermedades, se supondrá que estas poblaciones son las muestras de casos y controles para un fenotipo de enfermedad particular.

La estructura de espacio métrico garantiza la representación de los SNPs en un espacio euclidiano, haciendo uso de la metodología de escalamiento multidimensional no métrico. Esto permite la visualización de conglomerados de SNPs, de los cuales se seleccionan sus centroides o etiquetas de clase. Ofreciendo una metodología adicional en el marco de los estudios de asociación del genoma completo (GWAS).

### 3.1. Estructura de datos

Las muestras de ADN para el HapMap vienen de un total de 270 personas. El pueblo Yoruba de Ibadan, Nigeria, proporciona 30 conjuntos de muestras de dos padres y un hijo adulto (a cada uno se le llama un trío). En Japón, 45 individuos no relacionados de la zona de Tokio proporcionaron muestras. En China, 45 individuos no relacionados de Beijing proporcionaron muestras.

mTreinta tríos EE.UU. proporcionaron muestras, que fueron recogidos de los residentes de EE.UU. de ascendencia del norte de Europa en 1980 y el oeste por el Centre d'Etude du Polymorphisme Humain (CEPH) [9]. Según esta distribución se forman 3 poblaciones de estudio cada una con 90 personas para la muestra de haplotipos. Los detalles de las tablas de haplotipos se muestran con siglas de 3 letras utilizadas en los nombres de los archivos:

**CEU:** Residentes de Utah con ancestros del norte y Europa occidental.

**CHB:** Población China perteneciente a la etnia Han de Beijing.

**JPT:** Población japonesa residente en Tokio.

**YRI:** Yoruba en Ibadan, Nigeria (África Occidental).

**JPT + CHB:** Panel de Asia combinado.

Cada uno de las tablas contiene campos relevantes:

**Rs\_strand:** expresa la orientación de la hebra de ADN en la que se ubican los alelos, por tanto, (+) significa que los alelos se encuentra en orientación relativa a la referencia del genoma humano y (-) en el reverso.

**rs#:** representa la referencia del SNP.

**SNPalleles:** representan los alelos presentes en el SNP.

**Chrom:** identificador de cromosoma.

**Pos:** posición física del SNP, dentro del cromosoma.

**id Person:** 90 campos pertenecientes a cada id de persona, con la información de los alelos homocigotos y heterocigotos, con la posibilidad de la inexistencia de alelos dentro del SNP.

## 3.2. Tratamiento de datos y métodos

Usualmente un SNP es representado por un vector con coordenadas en  $\mathbb{Z}_3 = \{0, 1, 2\}$ , donde 0 está asociado al homocigoto con menor alelo, 1 está asociado al homocigoto con mayor alelo y 2 está asociado al alelo heterocigoto [12]. Otra forma de representar a un SNP como se verá más adelante, es formando una cadena por concatenación de las coordenadas en  $\mathbb{Z}_3$  del SNP.

De las tablas proporcionadas por el proyecto HapMap, se tomó el cromosoma 1 de dos poblaciones objetivo CEU (Caucásicas) y YRI (Yoruba) y se realizó el siguiente proceso para la formación y almacenamiento de los SNPs informativos.

Cabe anotar que estos SNPs  $s \in \mathbb{Z}_3^{90}$ , donde 90 es el tamaño de la muestra de personas. Los SNPs informativos sin elementos faltantes y no repetidos, forman un total de 108,619 SNPs para CEU y 130,052 SNPs para YRI.

Con el objetivo de comparar las poblaciones y encontrar asociaciones, se dividen las tablas  $T_p^*$  con  $p \in \{CEU, YRI\}$  en 3 partes suponiendo que cada bloque representa un fenotipo de interés.

Sean  $E = \{D, G, H\}$  el conjunto de índices, Diabetes tipo II, Gallstone e Hipertensión respectivamente, para efectos del ejemplo.  $CEU_i$  y  $YRI_i$  son los bloques formados a partir de  $T_p^*$  donde  $i \in E$ .

Se define  $\mapsto$  como el operador de inserción en una tabla.

---

**Procedimiento 1** Construcción de la tabla de SNPs informativos
 

---

**Entrada:**  $T_p$  tabla original de SNPs asociados a la población  $p$

**Salida:**  $T_p^{NN}$  tabla de SNPs con información incompleta o inexistente

$\overline{T}_p$  tabla de SNPs con información relevante

$T_p^*$  tabla de SNPs informativos con codificaciones diferentes

- 1: **para todo**  $s \in T_p$  **hacer**
  - 2:   **si** el SNP no contiene información de alelo
  - 3:   **si**  $NN \in s$  **entonces**  
        $s \mapsto T_p^{NN}$
  - 4:   **si no**  
        $s \mapsto \overline{T}_p$
  - 5:   **fin si**
  - 6: **fin para**
  - 7: **devolver**  $T_p^{NN}, \overline{T}_p$
  - 8: **para todo**  $s \in \overline{T}_p$  **hacer**
  - 9:   **se realiza** conteo de alelos, para su representación en  $\mathbb{Z}_3$   
        $s \xrightarrow{f_{\mathbb{Z}_3}} s^*$   
        $s^* \mapsto T_p^*$
  - 10: **fin para**
  - 11: **devolver**  $T_p^*$
- 

### 3.2.1. Métricas sobre $\mathbb{Z}_3^n$

Consideremos el conjunto  $\mathbb{Z}_3^n = \{0, 1, 2\}^n$ , de todas la  $n$ -tuplas de elementos en  $\mathbb{Z}_3$ .

Sobre este conjunto definimos 3 métricas:

1. Distancia  $L_1$  o métrica del taxista,  $\mathbf{d}_T$ .
2. Distancia de Hamming,  $\mathbf{d}_H$ .
3. Distancia de Levenshtein,  $\mathbf{d}_L$ .

Vamos a probar que bajo estas métricas el espacio  $(\mathbb{Z}_3^n, \mathbf{d}_i)$  es un espacio métrico para  $i \in \{H, T, L\}$ .

#### $(\mathbb{Z}_3^n, \mathbf{d}_T)$ es espacio métrico

**Definición 1.** La distancia  $L_1$ , denotada por  $\mathbf{d}_T$ , entre dos elementos  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$  es la suma de las longitudes de las proyecciones del segmento de línea entre los puntos sobre el sistema de ejes coordenados. Mas formalmente,

$$\mathbf{d}_T(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i| \quad \text{donde } x_i, y_i \in \mathbb{Z}_2$$

**Ejemplo 1.** Sea  $n = 5$ ,  $x = (0, 0, 1, 1, 2)$ ,  $y = (1, 1, 0, 0, 2)$ , luego  $\mathbf{d}_T(x, y) = 4$   
 Sea  $n = 3$ ,  $x = (2, 0, 1)$ ,  $y = (0, 1, 0)$ , luego  $\mathbf{d}_T(x, y) = 4$

**Teorema 1.** El par  $(\mathbb{Z}_3^n, \mathbf{d}_T)$  es un espacio métrico.

Para probar que la distancia  $L_1$ ,  $\mathbf{d}_T(\mathbf{x}, \mathbf{y})$  es una métrica sobre  $\mathbb{Z}_3^n$ , se debe cumplir:

1.  $\mathbf{d}_T(\mathbf{x}, \mathbf{y}) \geq 0 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$
2.  $\mathbf{d}_T(\mathbf{x}, \mathbf{y}) = 0$  si y solo si  $\mathbf{x} = \mathbf{y}$ ,  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$
3.  $\mathbf{d}_T(\mathbf{x}, \mathbf{y}) = \mathbf{d}_T(\mathbf{y}, \mathbf{x}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$
4.  $\mathbf{d}_T(\mathbf{x}, \mathbf{y}) + \mathbf{d}_T(\mathbf{y}, \mathbf{z}) \geq \mathbf{d}_T(\mathbf{x}, \mathbf{z}) \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{Z}_3^n$

*Demostración.*

□

1. Sean  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$ , por tanto  $x_i, y_i \in \mathbb{Z}_3 \quad \forall i = 1, \dots, n$ , de esto se sigue que  $|x_i - y_i| \in \mathbb{Z}_3$ , por tanto  $\mathbf{d}_T(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i| \geq 0$ .
2. Sean  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$ , tales que  $\mathbf{x} = \mathbf{y}$ , por tanto  $x_i = y_i \quad \forall i = 1, \dots, n$ , de esto se sigue que  $|x_i - y_i| = 0 \quad \forall i$  luego tenemos que  $\sum_{i=1}^n |x_i - y_i| = \mathbf{d}_T(\mathbf{x}, \mathbf{y}) = 0$ .
3. Sean  $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_3^n$

$$\mathbf{d}_T(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i| = \sum_{i=1}^n |y_i - x_i| = \mathbf{d}_T(\mathbf{y}, \mathbf{x})$$

4. Sean  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{Z}_3^n$

$$\begin{aligned} \mathbf{d}_T(\mathbf{x}, \mathbf{y}) + \mathbf{d}_T(\mathbf{y}, \mathbf{z}) &= \sum_{i=1}^n |x_i - y_i| + |y_i - z_i| \quad \text{por desigualdad triangular tenemos} \\ &\geq \sum_{i=1}^n |x_i - z_i| = \mathbf{d}_T(\mathbf{x}, \mathbf{z}) \quad |x_i - z_i| \leq |x_i - y_i| + |y_i - z_i| \quad \forall i = 1, \dots, n \end{aligned}$$

### Distancias de edición

En esta sección se considera el conjunto  $\overline{\mathbb{Z}_3^n}$  formado por todas las cadenas de longitud  $n$  con elementos en  $\mathbb{Z}_3$ . La distancia de edición entre dos cadenas es el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra. Actualmente los correctores de ortografía utilizan algoritmos basados en distancias de edición para detectar errores, corregir y proponer sugerencias de cadenas de caracteres. Entre los códigos más populares se encuentran el código de Hamming y el código de Levenshtein.

**Definición 2.** La distancia de **Hamming** denotada por  $\mathbf{d}_H$ , entre dos elementos  $\mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$  es el número de caracteres en los cuales difieren  $\mathbf{x}$  e  $\mathbf{y}$

**Teorema 2.** El par  $(\overline{\mathbb{Z}_3^n}, \mathbf{d}_H)$  es un espacio métrico.

Para probar que la distancia  $H$ ,  $\mathbf{d}_H(\mathbf{x}, \mathbf{y})$  es una métrica sobre  $\overline{\mathbb{Z}_3^n}$ , se debe cumplir:

1.  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) \geq 0$  y  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) = 0$  si y solo si  $\mathbf{x} = \mathbf{y}$ ,  $\forall \mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$
2.  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) = \mathbf{d}_H(\mathbf{y}, \mathbf{x}) \forall \mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$
3.  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) + \mathbf{d}_H(\mathbf{y}, \mathbf{z}) \geq \mathbf{d}_H(\mathbf{x}, \mathbf{z}) \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \overline{\mathbb{Z}_3^n}$

*Demostración.* □

1.  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) = 0$  si y solo si  $\mathbf{x}, \mathbf{y}$  coinciden en todas los caracteres y esto ocurre cuando  $\mathbf{x} = \mathbf{y}$ .
2. El número de coordenadas en las cuales  $\mathbf{x}$  difiere de  $\mathbf{y}$  es igual al número de caracteres en las cuales  $\mathbf{y}$  difiere de  $\mathbf{x}$ . Por tanto  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) = \mathbf{d}_H(\mathbf{y}, \mathbf{x})$ .
3.  $\mathbf{d}_H(\mathbf{x}, \mathbf{y})$  es el número mínimo de caracteres que difieren  $\mathbf{x}$  e  $\mathbf{y}$ .  $\mathbf{d}_H(\mathbf{y}, \mathbf{z})$  es el número mínimo de caracteres que difieren  $\mathbf{y}$  e  $\mathbf{z}$ .

Por tanto  $\mathbf{d}_H(\mathbf{x}, \mathbf{y}) + \mathbf{d}_H(\mathbf{y}, \mathbf{z})$  son los cambios posibles de caracteres que se pueden hacer para que  $\mathbf{x} = \mathbf{z}$ . Por tanto

$$\mathbf{d}_H(\mathbf{x}, \mathbf{y}) + \mathbf{d}_H(\mathbf{y}, \mathbf{z}) \geq \mathbf{d}_H(\mathbf{x}, \mathbf{z})$$

ya que  $\mathbf{d}_H(\mathbf{x}, \mathbf{z})$  son los mínimos cambios de caracteres que deben hacer para que  $\mathbf{x} = \mathbf{z}$ .

La distancia **Levenshtein**, fue creada e implementada por Vladimir Levenshtein a mediados del siglo XX, con el propósito de medir la diferencia entre dos secuencias de símbolos [15]. Para poder ver que esta distancia es una métrica sobre  $\overline{\mathbb{Z}_3^n}$ , es necesario definir el concepto de similitud entre secuencias [3].

**Definición 3.** Sean  $\mathbf{x} = x_1.x_2 \dots x_n$   $\mathbf{y} = y_1.y_2 \dots y_m$  con  $m \leq n$ , dos elementos de  $\overline{\mathbb{Z}_3^n}$ ,  $\mathbf{y}$  es una **subsecuencia** de  $\mathbf{x}$ , denotado como  $\mathbf{y} \subset \mathbf{x}$ , si existe un conjunto de índices  $\{i_1, i_2, \dots, i_m\}$  en  $\mathbf{x}$ , con cada  $1 \leq i_k \leq n$  y  $1 \leq k \leq m$ , tales que  $i_1 < i_2 < \dots < i_m$  y que  $\mathbf{y} = x_{i_1}.x_{i_2} \dots x_{i_m}$ .

**Definición 4.** Una **subsecuencia común** para las secuencias  $\mathbf{x}_a$  y  $\mathbf{x}_b$  denotado por  $\mathbf{y} \subset (\mathbf{x}_a, \mathbf{x}_b)$ , si  $\mathbf{y} \subset \mathbf{x}_a$  y  $\mathbf{y} \subset \mathbf{x}_b$ .

La similitud entre dos secuencias  $x, y \in \overline{\mathbb{Z}_3^n}$ , denotada por  $S(\mathbf{x}, \mathbf{y})$  está dada por:

$$S(\mathbf{x}, \mathbf{y}) = \text{máx} \{ \|\mathbf{z}\| \mid \mathbf{z} \subset (\mathbf{x}, \mathbf{y}) \text{ con } \mathbf{z} \in \overline{\mathbb{Z}_3^n} \}$$

donde  $\|\mathbf{z}\|$  indica la longitud de la secuencia  $\mathbf{z}$ , es decir, la cantidad de caracteres que contiene.

Se puede observar que  $S(\mathbf{x}, \mathbf{y}) = 0$ , cuando  $\mathbf{x}$  no tiene caracteres comunes con  $\mathbf{y}$ , esto es, no existe subsecuencia común. Por otro lado,  $S(\mathbf{x}, \mathbf{y}) = n$ , cuando  $\mathbf{x} \subset \mathbf{y}$  o  $\mathbf{y} \subset \mathbf{x}$ . Por tanto

$$0 \leq S(\mathbf{x}, \mathbf{y}) \leq n$$

Determinar la similitud de dos secuencias, se convierte entonces, en encontrar el tamaño de la **subsecuencia común más larga** entre las secuencias  $\mathbf{x}$  y  $\mathbf{y}$ .

La distancia de Levenshtein entre dos secuencias  $\mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$ , está definida como:

$$\mathbf{d}_L(\mathbf{x}, \mathbf{y}) = n - S(\mathbf{x}, \mathbf{y})$$

donde  $S(\mathbf{x}, \mathbf{y})$  es la similitud entre las secuencias  $\mathbf{x}$  y  $\mathbf{y}$ . Los límites de esta distancia se logran, por un lado cuando la similitud entre las secuencias comparadas es nula, y en el otro extremo, cuando la similitud entre las secuencias comparadas es máxima. Cuando la similitud es nula (secuencias sin caracteres comunes), la distancia es  $n$ . Cuando la similitud es máxima (se comparan secuencias iguales), la distancia es 0.

$$0 \leq \mathbf{d}_L(\mathbf{x}, \mathbf{y}) \leq n$$

La idea general de esta distancia, es que, dos secuencias distan entre sí tanto como caracteres se deban borrar y símbolos se deban agregar, para hacer iguales ambas secuencias. De modo que el límite máximo de esta distancia se debe leer como: *se deben borrar todos los  $n$  caracteres de  $\mathbf{x}$  y agregar todos los caracteres de  $\mathbf{y}$*  [3].

**Teorema 3.** *El par  $(\overline{\mathbb{Z}_3^n}, \mathbf{d}_L)$  es un espacio métrico.*

Para probar que la distancia  $\mathbf{d}_L$  es una métrica sobre  $\overline{\mathbb{Z}_3^n}$ , se debe cumplir las cuatro condiciones del teorema 1.

*Demostración.* □

1. Se sigue por definición, ya que,  $0 \leq \mathbf{d}_L(\mathbf{x}, \mathbf{y}) \leq n$ .
2. Sean  $\mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$ ,
  - Supongamos  $\mathbf{x} = \mathbf{y}$ , por tanto  $S(\mathbf{x}, \mathbf{y}) = n$ , luego  $\mathbf{d}_L(\mathbf{x}, \mathbf{y}) = n - S(\mathbf{x}, \mathbf{y}) = n - n = 0$
  - Supongamos  $\mathbf{d}_L(\mathbf{x}, \mathbf{y}) = 0$ , luego  $S(\mathbf{x}, \mathbf{y}) = n$ , pero esto ocurre en el caso  $\mathbf{x} = \mathbf{y}$ .

3. Sea  $\mathbf{k} \subset (\mathbf{x}, \mathbf{y})$ , una subsecuencia común de  $\mathbf{x}$  y  $\mathbf{y}$ , entonces  $\|\mathbf{k}\| = S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$ , por tanto

$$\mathbf{d}_L(\mathbf{x}, \mathbf{y}) = n - S(\mathbf{x}, \mathbf{y}) = n - S(\mathbf{y}, \mathbf{x}) = \mathbf{d}_L(\mathbf{y}, \mathbf{x})$$

4.  $\mathbf{d}_L(\mathbf{x}, \mathbf{y}) + \mathbf{d}_L(\mathbf{y}, \mathbf{z}) \leq \mathbf{d}_L(\mathbf{x}, \mathbf{z})$  por definición

$$\begin{aligned} n - S(\mathbf{x}, \mathbf{y}) + n - S(\mathbf{y}, \mathbf{z}) &\geq n - S(\mathbf{x}, \mathbf{z}) \\ n - S(\mathbf{y}, \mathbf{z}) - S(\mathbf{x}, \mathbf{y}) + S(\mathbf{x}, \mathbf{z}) &\geq 0 \end{aligned}$$

Se probará esto último.

Sea  $\alpha \subset (\mathbf{x}, \mathbf{y})$  y  $\beta \subset (\mathbf{y}, \mathbf{z})$ , por tanto  $\|\alpha\| = S(\mathbf{x}, \mathbf{y})$  y  $\|\beta\| = S(\mathbf{y}, \mathbf{z})$ , respectivamente. Además  $\|\alpha\| \leq \|\mathbf{y}\|$  y  $\|\beta\| \leq \|\mathbf{y}\|$ .

Como  $\alpha$  y  $\beta$  comparten caracteres de  $\mathbf{y}$ , existe  $\delta \subset (\alpha, \beta) \subset \mathbf{y}$ , luego  $\|\delta\| = \min\{\|\alpha\|, \|\beta\|\}$ . De modo que,  $\delta \subset (\mathbf{x}, \mathbf{y})$ , porque  $\alpha \subset \mathbf{x}$  y  $\beta \subset \mathbf{y}$ , también  $\delta \subset (\mathbf{y}, \mathbf{z})$ , aunque no necesariamente la de mayor longitud, luego  $\|\delta\| \leq S(\mathbf{x}, \mathbf{z})$ .

Entonces,  $\|\alpha\|$  representa la cantidad de caracteres que comparten  $\mathbf{x}$  y  $\mathbf{y}$ ,  $\|\beta\|$  es la cantidad de caracteres que comparten  $\mathbf{y}$  y  $\mathbf{z}$ , y  $\|\delta\|$  es la cantidad de caracteres que pertenecen a  $\mathbf{y}$  solamente.

Por tanto,  $\|\alpha\| + \|\beta\| - \|\delta\|$  es la cantidad de caracteres de  $\mathbf{y}$  que son compartidos con  $\mathbf{x}$  y  $\mathbf{z}$ , significa que  $\|\mathbf{y}\| \geq \|\alpha\| + \|\beta\| - \|\delta\|$ , en consecuencia

$$\|\mathbf{y}\| - \|\alpha\| - \|\beta\| + \|\delta\| \geq 0$$

Como  $\|\delta\| \leq S(\mathbf{x}, \mathbf{z})$ ,  $\|\alpha\| = S(\mathbf{x}, \mathbf{y})$  y  $\|\beta\| = S(\mathbf{y}, \mathbf{z})$ , tenemos

$$\begin{aligned} \|\mathbf{y}\| - S(\mathbf{y}, \mathbf{z}) - S(\mathbf{x}, \mathbf{y}) + S(\mathbf{x}, \mathbf{z}) &\geq 0 \\ n - S(\mathbf{y}, \mathbf{z}) - S(\mathbf{x}, \mathbf{y}) + S(\mathbf{x}, \mathbf{z}) &\geq 0 \end{aligned}$$

Bajo estos resultados se tiene la siguiente desigualdad.

$$\mathbf{d}_L(\mathbf{x}, \mathbf{y}) \leq \mathbf{d}_H(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \overline{\mathbb{Z}_3^n}$$

### 3.2.2. Algoritmos de Implementación

Se presentan los pseudocódigos para la implementación de las distancias  $L_1$ , distancias de edición.

---

**Procedimiento 2**  $d_T : \mathbb{Z}_3^n \times \mathbb{Z}_3^n \rightarrow \mathbb{N}$  Cálculo de la distancia  $L_1$ 


---

**Entrada:**  $s^{(1)}, s^{(2)} \in \mathbb{Z}_3^n$ **Salida:**  $d_T(s^{(1)}, s^{(2)})$ 

- 1:  $s = 0$  //se inicializa sumatoria acumulada
  - 2: **para**  $i = 1 : n$  **hacer**  
 $s = s + |s_i^{(1)} - s_i^{(2)}|$
  - 3: **fin para**
  - 4: **devolver**  $d_T(s^{(1)}, s^{(2)}) = s$
- 

---

**Procedimiento 3**  $d_H : \overline{\mathbb{Z}_3^n} \times \overline{\mathbb{Z}_3^n} \rightarrow \mathbb{N}$  Cálculo de la distancia de *Hamming*


---

**Entrada:**  $s^{(1)}, s^{(2)} \in \overline{\mathbb{Z}_3^n}$ **Salida:**  $d_H(s^{(1)}, s^{(2)})$ 

- 1:  $c = 0$  //se inicializa variable de costo
  - 2: **para**  $i = 1 : n$  **hacer**
  - 3: **si**  $s_i^{(1)} \neq s_i^{(2)}$  **entonces**  
 $c = c + 1$
  - 4: **fin si**
  - 5: **fin para**
  - 6: **devolver**  $d_H(s^{(1)}, s^{(2)}) = c$
- 

---

**Procedimiento 4**  $d_L : \overline{\mathbb{Z}_3^n} \times \overline{\mathbb{Z}_3^n} \rightarrow \mathbb{N}$  Cálculo de la distancia de *Levenshtein*


---

**Entrada:**  $s^{(1)}, s^{(2)} \in \overline{\mathbb{Z}_3^n}$ **Salida:**  $d_H(s^{(1)}, s^{(2)})$ 

- 1:  $d = \mathbf{0}_{(n+1) \times (n+1)}$  //matriz nula  $n + 1 \times n + 1$
  - 2:  $c = 0$  //se inicializa variable de costo
  - 3: **para**  $i = 0 : n$  **hacer**  
 $d_{i,0} = i$
  - 4: **fin para**
  - 5: **para**  $j = 0 : n$  **hacer**  
 $d_{0,j} = j$
  - 6: **fin para**
  - 7: **para**  $i = 1 : n$  **hacer**
  - 8: **para**  $j = 1 : n$  **hacer**
  - 9: **compara** el caracter  $i$  de  $s^{(1)}$  con cada uno de los caracter de  $s^{(2)}$  cuando encuentra que son iguales no hay ningun ‘‘costo’’ en una de las operaciones definidas para Levenshtein en caso contrario si existe ‘‘costo’’
  - 10: **si**  $s_{i-1}^{(1)} \equiv s_{j-1}^{(2)}$  **entonces**  
 $c = 0$
  - 11: **si no**  
 $c = 1$
  - 12: **fin si**
  - 13: **se hacen** las operaciones de borrar, insertar y sustituir  
 $d_{i,j} = \min(d_{i-1,j} + 1, d_{i,j-1} + 1, d_{i-1,j-1} + c)$
  - 14: **fin para**
  - 15: **fin para**
  - 16: **devolver**  $d_H(s^{(1)}, s^{(2)}) = d_{n+1,n+1}$
-

### 3.2.3. Comparación de poblaciones

Las distancias de edición han sido utilizadas frecuentemente en problemas de bioinformática para el análisis de secuencias de ADN, la idea central del uso de distancias de edición, es encontrar similitudes entre secuencias partiendo del supuesto de que secuencias similares deben tener funciones similares dentro del genoma, en particular, dentro de los bloques de SNPs.

Como se describe en el procedimiento 1, se obtiene la tabla  $T_p^*$  compuesta de la muestra de SNPs informativos para  $p \in \{CEU, YRI\}$  y como se menciona en la pagina 13, se obtienen tres pares de muestras que representan los fenotipos de estudio.

Para el proceso de comparación de poblaciones o bloques de SNPs (fenotipos), se toma como población de control los bloques pertenecientes a la población YRI y como población enferma los bloques pertenecientes a la población CEU, la escogencia es debido al hecho de que  $|T_{YRI}^*| \geq |T_{CEU}^*|$ . La estructura de las tablas de SNPs para desarrollar el filtro de correlación se muestra en la tabla **3-2**.

**Tabla 3-2:** Estructura de SNPs casos-control para los fenotipos *DII*, *GLS*, *HIP*.

Fen	SNP	Casos	SNP $\mathbb{Z}_3$	SNP	Control	SNP $\mathbb{Z}_3$
DII	rs2977656	YRI	111111111 ... 1111	rs2341354	CEU	112212012 ... 1221
DII	rs2286139	YRI	122111112 ... 1221	rs2465126	CEU	111111111 ... 1211
DII	rs12562034	YRI	111112211 ... 1111	rs2465136	CEU	112112212 ... 1112
⋮	⋮	⋮	⋮	⋮	⋮	⋮
DII	rs2905035	YRI	122111112 ... 1221	rs9442371	CEU	112212212 ... 1212
GLS	rs819976	YRI	101001111 ... 1001	rs6670776	CEU	010122112 ... 1201
GLS	rs9439462	YRI	100011112 ... 1111	rs6688000	CEU	110022111 ... 1201
GLS	rs1987191	YRI	110022101 ... 1221	rs745910	CEU	121220012 ... 1222
⋮	⋮	⋮	⋮	⋮	⋮	⋮
GLS	rs1571150	YRI	122111000 ... 1201	rs2474460	CEU	110212212 ... 1200
HIP	rs1964052	YRI	101111111 ... 2222	rs12138757	CEU	010000112 ... 1111
HIP	rs164146	YRI	222211110 ... 2112	rs11808074	CEU	110022000 ... 1200
HIP	rs164147	YRI	220022111 ... 1020	rs164424	CEU	100000012 ... 1000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
HIP	rs164148	YRI	111111002 ... 2011	rs16861748	CEU	112012212 ... 1000

Se define una medida de correlación sobre el espacio  $\overline{\mathbb{Z}_3^{90}}$ , utilizando como input la distancia de edición de Levenshtein.

$$\mathbf{d}_L : [0, 90] \rightarrow \rho : [0, 1]$$

$$x \mapsto 1 - \frac{x}{90}$$

Esta medida de correlación define una medida de similaridad entre la población de control *YRI* y la población enferma *CEU*.

Sea  $\rho^*$  un nivel objetivo de correlación. Se puede generar una nueva tabla de SNPs de la población enferma  $T_{\rho(CEU)}$ , de tal forma  $\rho(s_{CEU}, s_{YRI}) \geq \rho^*$ . La construcción de  $T_{\rho(CEU)}$  se describe en el siguiente procedimiento.

---

**Procedimiento 5** Construcción de la tabla de SNPs  $T_{\rho(CEU)}$

---

**Entrada:**  $T_{CEU}^*$  tabla de SNPs informativos para la población *CEU*

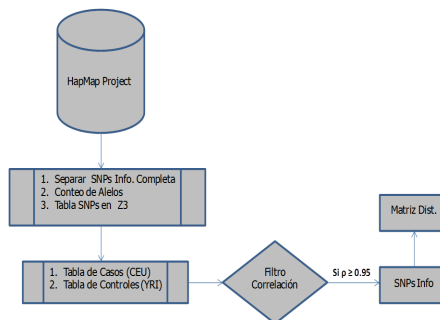
$T_{YRI}^*$  tabla de SNPs informativos para la población *YRI*

$\rho^*$  nivel de correlación objetivo

**Salida:**  $T_{\rho(CEU)}$  tabla de SNPs,  $s_{CEU}$  correlacionados a un nivel  $\rho^*$  con  $s_{YRI}$

- 1: **para todo**  $s_e \in T_{CEU}^*$  **hacer**
  - 2:   **mientras**  $s_c \in T_{YRI}^*$  **hacer**
  - 3:     //se calcula la correlación entre  $s_e$  y  $s_c$
  - 4:     **si**  $\rho(s_e, s_c) \geq \rho^*$  **entonces**
  - 5:        $s_e \mapsto T_{\rho(CEU)}$
  - 6:     **fin si**
  - 7:   **fin mientras**
  - 8: **devolver**  $T_{\rho(CEU)}$
- 

A partir de la tabla  $T_{\rho(CEU)}$  se construye la matriz de distancias que es el input para el análisis de escalamiento multidimensional de la muestra de SNPs que representa a la población enferma y que esta correlacionada a un nivel  $\rho^*$  con la población de control.



**Figura 3-1:** Proceso de filtros para la obtención de SNPs informativos.

# 4 Metodologías de Representación y Conglomeración de SNPs

## 4.1. Escalamiento Multidimensional en el espacio de SNPs

El Escalamiento Multidimensional (*Multidimensional Scaling, MDS*) es una técnica multivariada de dependencia de los datos que trata de representar en un espacio geométrico de pocas dimensiones las proximidades existentes en un conjunto de objetos. Esta técnica parte de las distancias o disimilaridades establecidas entre un conjunto de  $n$  objetos y conduce a la construcción o representación de estos en un espacio métrico, generalmente euclidiano. Según [5] los principales propósitos del escalamiento multidimensional son los siguientes

- I. Es un método que representa las disimilaridades de los datos como distancias en un espacio euclidiano para hacer que estos sean accesibles a la inspección visual y exploración.
- II. Es una técnica que permite verificar si las diferencias, que distingue a unos objetos de otros, se refleja en la representación conseguida.
- III. Es una aproximación analítica a los datos que permite descubrir las dimensiones relevantes presentes en las disimilaridades.

Supongamos que los datos de partida constituyen un conjunto de  $n$  individuos de los cuales se tiene información sobre su disimilaridad o distancia entre ellos. Se puede definir una matriz de entrada  $\Delta = \delta_{ij}$  cuyos elementos son el resultado de generar las proximidades para cualquier par de individuos  $(i, j)$ . Esta matriz es simétrica, luego  $\delta_{ij} = \delta_{ji}$ . Según [16], el objetivo principal del MDS es encontrar una matriz de salida de coordenadas entre los individuos  $X = x_{ik}$  en  $\mathbb{R}^m$ , donde las filas son los individuos  $i = 1, \dots, n$  y las columnas son las dimensiones del espacio  $k = 1, \dots, m$ .

Los principales tipos de escalamiento multidimensional son los siguientes [5].

**EMD-M:** El *Escalamiento Multidimensional Métrico* o clásico, asume que las disimilaridades son de tipo euclidiano y que se puede establecer una relación de tipo lineal con las distancias en el espacio de representación. Por tanto, se busca una función lineal  $f$  tal

que,  $d_{ij} \approx f(\delta_{ij}) = \alpha + \beta\delta_{ij}$ . La estimación de los parámetros  $\alpha$  y  $\beta$ , se obtienen por método de mínimos cuadrados. La función no necesariamente debe ser lineal, también se pueden ajustar otras funciones paramétricas que cumplan con la condición de ser continuas y monótonas (creciente o decreciente).

**EMD-NM:** El *Escalamiento Multidimensional No Métrico* u ordinal, no asume una relación paramétrica entre la disimilaridades y las distancias, es usado cuando las transformaciones de las disimilaridades no conservan la magnitud de las variables pero mantienen las propiedades de orden o monotonía, por tanto,

$$\delta_{i'i'} \leq \delta_{j'j'} \text{ entonces } f(\delta_{i'i'}) \leq f(\delta_{j'j'})$$

Para todos los elementos  $0 \leq i, i', j', j' \leq n$ . El desarrollo de este método es debido a Shepard en [20] quien demostró que es posible obtener soluciones métricas asumiendo únicamente una relación ordinal entre proximidades y distancias. Posteriormente Kruskal en [14] mejoró el modelo propuesto inicialmente Shepard.

En el capítulo 3 consideramos dos posibles representaciones del espacio de SNPs.

- I. El conjunto de SNPs pertenece al espacio métrico  $(\mathbb{Z}_3^n, \mathbf{d}_T)$  donde  $n = 90$  y  $\mathbf{d}_T$  es la métrica  $L_1$ .
- II. El conjunto de SNPs pertenece al espacio métrico  $(\overline{\mathbb{Z}}_3^n, \mathbf{d}_I)$  donde  $n = 90$  e  $I = \{H, L\}$  son las distancias de Hamming y Levenshtein, respectivamente.

Dado que el espacio de SNPs no es espacio vectorial y las distancias (disimilaridades) definidas en este trabajo no son de tipo euclidiano, se hace factible el uso del método de *EMD-NM*. El algoritmo es presentado en el procedimiento 6.

Una estrategia gráfica para medir el ajuste del modelo consiste en disponer sobre el eje horizontal, las disimilaridades y sobre el eje vertical las distancias estimadas en cada proceso iterativo; la configuración final de este diagrama se conoce como *diagrama de Shepard* [5].

Para este método es necesario obtener un coeficiente que de información sobre la bondad del ajuste. Las distancias sobre el espacio de representación son función de las disimilitudes  $d_{ij} \approx f(\delta_{ij})$ , el proceso iterativo va generando nuevas transformaciones de los puntos en el espacio de representación. A las transformaciones de las disimilitudes por  $f$  se denomina *disparidades*, a partir de estas disparidades y las distancias estimadas se puede definir el error cuadrático como  $e_{ij}^2 = (f(\delta_{ij}) - d_{ij})^2$  [6].

Para medir qué tan buena es la representación que se genera en cada iteración, se calcula la estadística *Stress*, la cual Kruskal definió como

$$S = \frac{\sum_{i \neq j} e_{ij}^2}{\sum_{i \neq j} d_{ij}^2}$$

**Procedimiento 6** Descripción del método *EMD – NM*.**Entrada:**  $\Delta = (\delta_{ij})$ ,  $\Delta \in M_{n \times n}$  Matriz de disimilaridades $k$  Dimensión del espacio de representación

maxit Número máximo de iteraciones

 $\epsilon$  Criterio de convergencia**Salida:**  $X \in \mathbb{R}_{n \times k}$  Matriz de representación $s$  Estadístico Stress1:  $\Delta \mapsto Rk$  Matriz de rangos, desde 1 hasta  $\frac{n(n-1)}{2}$ 2:  $X_0 \in \mathbb{R}_{n \times k}$  Matriz de coordenadas aleatorias3:  $\hat{D}_0$  Distancias entre elementos de  $X_0$ 4:  $s_0$  Cálculo de Stress5: **si**  $|s_0| < \epsilon$  **entonces**    **salir**6: **si no**7:   **para**  $i = 1, \dots, \text{maxit}$  **hacer**8:      $X \in \mathbb{R}_{n \times k}$ 9:      $\hat{D}$ 10:      $s$  Cálculo de Stress11:     **si**  $|s - s_0| < \epsilon$  **entonces**      **salir**12:     **si no**13:        $s_0 = s$  Se actualiza el Stress14:     **fin si**15:   **fin para**16: **fin si**17: **devolver**  $s, X$

El *Stress* es la suma de cuadrados residuales normalizados, de tal forma que el rango de esta estadística es  $[0, 1]$ . El denominador de  $\mathcal{S}$  es un factor de escala, inferido desde los  $\delta$ 's [5]. El *Stress* se expresa frecuentemente en porcentaje, Kruskal en [14] sugiere las siguientes interpretaciones del *Stress*.

- I.  $\mathcal{S} = 0$  la configuración de los datos en el espacio de representación es *Excelente*.
- II.  $\mathcal{S} \in (0, 0,05]$  la configuración de los datos en el espacio de representación es *Buena*.
- III.  $\mathcal{S} \in (0,05, 0,1]$  la configuración de los datos en el espacio de representación es *Acceptable*.
- IV.  $\mathcal{S} > 0,1$  la configuración de los datos en el espacio de representación es *Pobre*.

La diferenciabilidad de  $\mathcal{S}$  es utilizado en el proceso iterativo que asemeja al método del *descenso más pendiente*, para tratar de encontrar el ajuste que produzca el mínimo valor de  $\mathcal{S}$  [5]. La salida del proceso iterativo es la matriz de representación, el valor del *Stress*  $\mathcal{S}$  y el número de iteraciones realizadas.

## 4.2. EMD-NM para la muestra de fenotipos

Para encontrar asociaciones entre enfermedades utilizando marcadores genéticos o SNPs es necesario definir dos poblaciones, SNPs pertenecientes a la población sana (SNPs controles) y SNPs pertenecientes a la población enferma (SNPs casos). Para obtener una muestra global que representen las poblaciones enferma y de control se utiliza el procedimiento 1 descrito en el capítulo 3. Se define como grupo de control al conjunto de SNPs pertenecientes a la población *YRI* y como grupo de casos al conjunto de SNPs pertenecientes a la población *CEU*. Que en términos del procedimiento 1 los SNPs informativos quedan almacenados en las tablas  $T_{CEU}^*$  (casos) y  $T_{YRI}^*$  (controles).

En nuestro caso existen tres fenotipos de interés, *D*: Diabetes tipo II, *G*: Enfermedad de Gallstone e *H*: Hipertensión. Para obtener las muestras *casos - controles* para cada fenotipo, se divide en tres partes las tablas  $T_{CEU}^*$  y  $T_{YRI}^*$  obteniéndose los tres pares de bloques  $CEU_i$  y  $YRI_i$  con  $i \in \{D, G, H\}$ , que representan las muestras de SNPs informativos *casos - controles* para cada uno de los fenotipos, como se mencionó en el capítulo 3.

Sobre el espacio de SNPs se define como métrica la *Distancia Levenshtein*  $\mathbf{d}_L$  y como se describe en la sección 3.2.3, se puede definir una medida de correlación a partir de la distancia de edición, que es útil para realizar el filtro descrito en el procedimiento 5. Utilizando como nivel objetivo  $\rho^* = 0,95$  se efectúa el filtro para cada uno de los tres pares de bloques de SNPs  $CEU_i$  y  $YRI_i$  con  $i \in \{D, G, H\}$ , estos SNPs quedan almacenados en las tablas  $T_{\rho(CEU_i)}$  con  $i \in \{D, G, H\}$  y representan el conjunto de SNPs de la población enferma (casos) que están correlacionados sobre un nivel  $\rho^*$  con la algún SNP de la población sana (control), esto es, son los SNPs de la población enferma que explican o aportan mayor

información a la población sana. Este razonamiento está basado en que los SNPs (variables en contexto estadístico clásico), con información correlacionada en ambos fenotipos, son los mas informativos para explicar la presencia de la enfermedad en relación con el fenotipo. El número de SNPs relevantes para el análisis de escalamiento multidimensional se muestra a continuación.

- $|T_{\rho(CEU_D)}| = 1310$  de 17761 originalmente.
- $|T_{\rho(CEU_G)}| = 1218$  de 17761 originalmente.
- $|T_{\rho(CEU_H)}| = 1443$  de 17762 originalmente.

A partir de estas tablas se calculan las matrices de distancias  $\Delta_{\rho(CEU_i)}$  con  $i \in \{D, G, H\}$  utilizando como métrica la *Distancia Levenshtein*  $\mathbf{d}_L$ . Estas matrices son el input del análisis de escalamiento multidimensional descrito en el procedimiento 6. Las matrices salida de coordenadas entre los SNPs  $X_{\rho(CEU_i)}$  en  $\mathbb{R}^m$  donde  $i \in \{D, G, H\}$  y  $m = \{2, 3\}$ , se obtuvieron a través del software libre R Development Core Team [18] y las librerías `NeatMap`, `ecodist` y `smacof` [16].

#### 4.2.1. EMD-NM para el fenotipo Diabetes tipo II

Para encontrar asociaciones entre las dos poblaciones *casos - controles* ligados al fenotipo Diabetes tipo II se realiza un cambio de espacio, esto es, conociendo la relación de cercanía o disimilaridad en el espacio SNPs se busca una distribución de estos SNPs en el espacio euclidiano  $\mathbb{R}^2$  y  $\mathbb{R}^3$ .

A partir de la tabla  $T_{\rho(CEU_D)}$  se construye la matriz de distancias  $\Delta_{\rho(CEU_D)}$  el cual es input para el *EMD - NM*. En la tabla 4-1 se observa que el valor del Stress disminuye a medida que la dimensión del espacio de representación aumenta y esto hace que el tiempo de ejecución y número de iteraciones aumente proporcionalmente.

**Tabla 4-1:** EMD-NM para el fenotipo Diabetes tipo II en  $R^2$  y  $R^3$ .

Num Dim	Stress NM	Num Iter
2	0,05563	226
3	0,02426	969

En la figura 4-1 se observa la distribución de SNPs en los espacios  $\mathbb{R}^2$  y  $\mathbb{R}^3$ , así como sus respectivos diagramas de Shepard asociados a estas distribuciones. Durante el proceso iterativo se generan distintas representaciones de SNPs en el espacio euclidiano, a su vez, se puede calcular las distancias entre ellos denotadas  $d_{ij}$ , el diagrama de Shepard es una dispersión de puntos entre disimilaridades  $\delta_{ij}$  (espacio de SNPs) y las distancias  $d_{ij} = f(\delta_{ij})$  (espacio euclidiano), lo puntos huecos son los pares de disimilaridades y distancias  $(\delta_{ij}, d_{ij})$

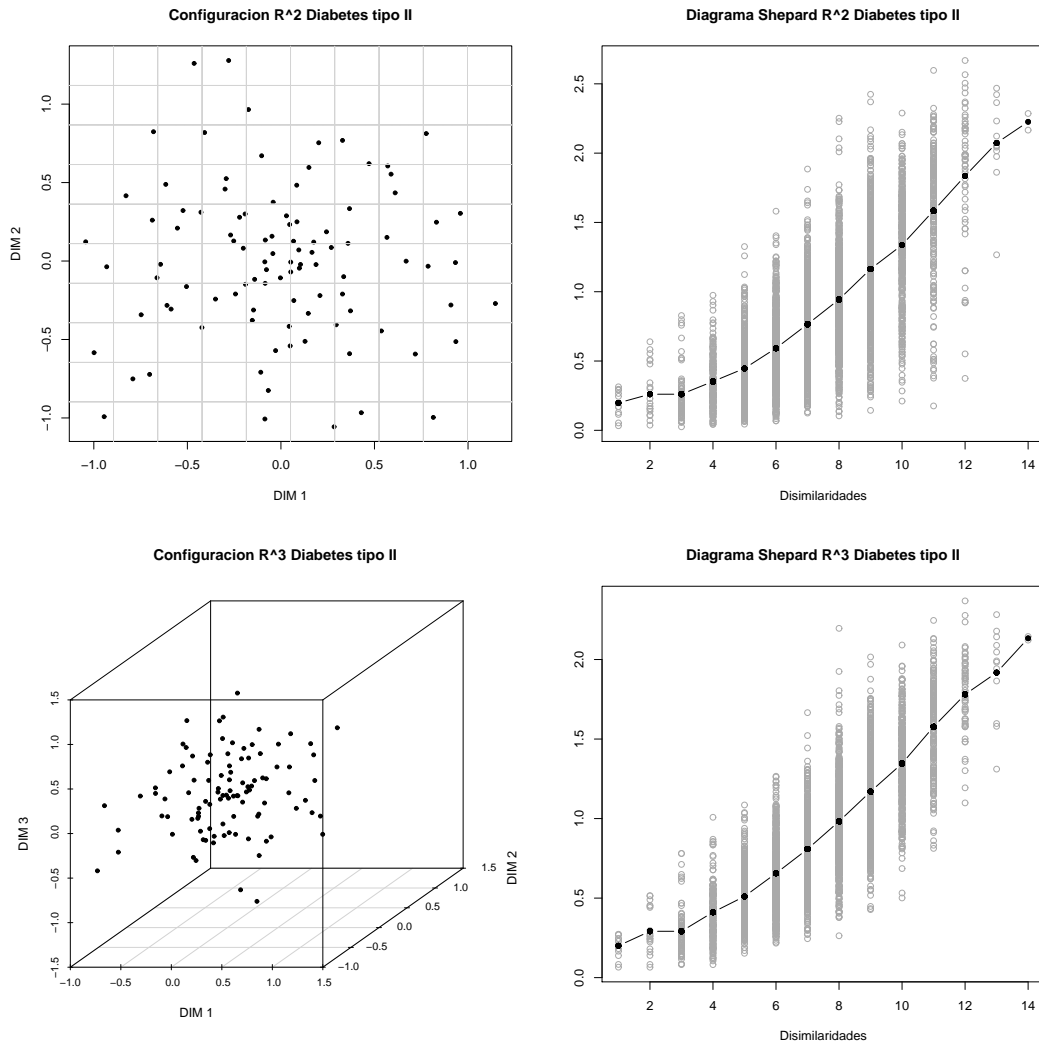


Figura 4-1: Configuración de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  y Diagramas de Shepard.

y los puntos llenos son los pares de disimilaridades y disparidades  $(\delta_{ij}, \hat{d}_{ij})$  donde,  $\hat{d}_{ij}$  es la distancia estimada luego de minimizar el *Stress*  $\mathcal{S}$ .

En la tabla 4-2 se muestra la representación de SNPs en los espacios  $\mathbb{R}^2$  y  $\mathbb{R}^3$ . La contribución al valor del *Stress* objetivo por cada SNP es mostrado en la tabla 4-3.

El procedimiento y salidas del método *EMD* – *NM* para los otros fenotipos es mostrado en el apéndice 8

**Tabla 4-2:** Configuración en  $R^2$  y  $R^3$  para el fenotipo Diabetes tipo II.

SNP	D1	D2	SNP	D1	D2	D3
rs6662245	-0,27936	1,27797	rs6662245	-0,59282	0,99021	0,42239
rs278849	-0,14054	-0,11657	rs278849	-0,13942	-0,13867	-0,02169
rs1539737	-0,82816	0,41588	rs1539737	-0,76376	0,30895	-0,39228
rs2032056	-0,42753	0,31097	rs2032056	-0,40929	0,21858	-0,31048
rs1166702	-0,19152	0,29987	rs1166702	-0,24399	0,26948	-0,09927
rs17101082	-0,61559	0,48859	rs17101082	-0,51095	0,39112	-0,44742
rs1780049	0,04904	-0,54094	rs1780049	0,05521	-0,53132	-0,27755
rs10873963	-0,29875	0,45856	rs10873963	-0,29722	0,45306	-0,21842
rs638335	-0,46399	1,25959	rs638335	-0,63335	1,05219	0,15398
rs17101412	-0,25229	0,12766	rs17101412	-0,07695	0,0793	-0,50233
rs6694656	-0,2926	0,52507	rs6694656	-0,40912	0,42913	0,24877
rs2025292	0,36677	-0,59051	rs2025292	0,52609	-0,25398	-0,4177
rs17101646	0,71792	-0,59274	rs17101646	0,36821	-0,62743	0,59399
rs7533564	-0,55344	0,20952	rs7533564	-0,29873	-0,06998	-0,68053
rs11809462	-0,26902	0,16584	rs11809462	-0,12829	-0,01856	-0,5405
rs11590908	-0,66253	-0,10677	rs11590908	-0,44647	0,01072	0,55563
rs1929984	-0,02856	-0,57088	rs1929984	0,04179	-0,59045	0,0112
rs1322931	-0,68752	0,26043	rs1322931	-0,59039	0,08175	-0,48886
rs12068437	-0,22109	0,27896	rs12068437	-0,2055	0,24294	-0,16644
rs1413357	-0,08581	-1,0067	rs1413357	-0,12191	-0,89493	0,34168

### 4.3. Análisis de conglomerados por agrupamiento

#### K-Means

El agrupamiento K-Means es un método para encontrar grupos y centros de grupo o *centroides* en un conjunto de datos. Se elige un número de centroides  $k$  para los grupos y el procedimiento iterativamente desplaza los centroides para reducir al mínimo la varianza total dentro del grupo [10].

Se empieza considerando un conjunto de datos  $x_1, \dots, x_N$  que consta de una muestra de  $N$  observaciones de una variable aleatoria  $X$  sobre un espacio euclidiano  $d$ -dimensional. El objetivo es dividir el conjunto de datos en un número  $K$  de grupos, donde este valor de  $K$  es dado. Intuitivamente un conglomerado es un grupo comprimido de datos los cuales las distancias entre puntos del grupo son pequeñas comparadas con las distancias a los puntos fuera de la agrupación.

Para esto, se introduce un conjunto de vectores  $d$ -dimensionales  $\{\mu_k\}_{k=1}^K$  asociado al  $k$ -ésimo grupo. Estos  $\{\mu_k\}_{k=1}^K$  son los centroides iniciales de los conglomerados. El objetivo entonces es encontrar una asignación de puntos de datos para cada grupo, así como un conjunto de vectores  $\{\mu_k^*\}_{k=1}^K$ , de tal manera que la suma de los cuadrados de las distancias de cada punto

**Tabla 4-3:** Stress por punto en  $R^2$  y  $R^3$  para el fenotipo Diabetes tipo II.

SNP	StPP $R^2$	SNP	StPP $R^3$
rs17106643	0,01177	rs17106643	0,00565
rs12068437	0,01197	rs278849	0,0058
rs12090854	0,01412	rs17101412	0,00589
rs278849	0,01498	rs12090854	0,00668
rs17106720	0,0153	rs17106696	0,00773
rs12121256	0,01698	rs12121256	0,00791
rs3904037	0,01732	rs2388956	0,00805
rs17106696	0,01759	rs4417015	0,00827
rs4417015	0,01777	rs12133546	0,00859
rs12133546	0,01822	rs11163325	0,00872
rs12740817	0,01856	rs356294	0,00891
rs12746595	0,01878	rs17106720	0,00921
rs11163325	0,01987	rs12068437	0,00923
rs273231	0,02008	rs273231	0,00929
rs1281601	0,02051	rs1929984	0,00992
rs1929984	0,02135	rs3904037	0,01052
rs17103834	0,02154	rs7412036	0,01087
rs1166702	0,02218	rs10874230	0,01093
rs7538529	0,02257	rs7538529	0,01133
rs7522218	0,02278	rs17403822	0,01197

al vector  $\mu_k^*$  más cercano, sea mínima.

Para cada punto  $x_n$ , se introduce un conjunto de variables indicadoras binarias  $r_{nk} \in \{0, 1\}$ , donde  $k = 1, \dots, K$  y  $n = 1, \dots, N$ , que determina la presencia o ausencia del punto  $x_n$  en el grupo  $k$ , por tanto, si un punto  $x_n$  pertenece al grupo  $k$  entonces  $r_{nk} = 1$  y  $r_{nj} = 0$  si  $j \neq k$ . Luego se define una función objetivo  $J$ , llamada *medida de distorsión* dada por,

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2 \quad (4-1)$$

El cual representa la suma de cuadrados de las distancias de cada punto al vector asignado  $\mu_k$ . Se deben encontrar valores de  $r_{nk}$  y  $\mu_k$  de tal forma que se minimice  $J$ . Se puede establecer un proceso iterativo en el que cada iteración consiste en dos etapas sucesivas correspondientes a las optimizaciones sucesivas con respecto a  $r_{nk}$  y  $\mu_k$ .

En primer lugar, se eligen aleatoriamente algunos valores iniciales de  $\mu_k$ . Luego, en la primera fase se minimiza  $J$  con respecto a  $r_{nk}$ , manteniendo  $\mu_k$  fijo. En la segunda fase se minimiza  $J$  con respecto a la  $\mu_k$ , manteniendo  $r_{nk}$  fijo. Esta optimización en dos etapas se repite hasta que haya convergencia. En cada etapa del proceso se actualizan  $r_{nk}$  y  $\mu_k$ .

La primera fase es la determinación de  $r_{nk}$ . Como  $J$  en 4-1 es una función lineal de  $r_{nk}$ , esta optimización se puede realizar fácilmente y dar una solución en forma cerrada.

$$\frac{\partial J}{\partial r_{nk}} = \sum_{n=1}^N \sum_{k=1}^K \|x_n - \mu_k\|^2$$

Los términos para diferentes  $n$  son independientes, por tanto,  $J$  se pueden optimizar por separado eligiendo de  $r_{nk} = 1$ , escogiendo  $k$  como el subíndice para el cual  $\|x_n - \mu_k\|^2$  es el mínimo  $\forall k = 1, \dots, K$ .

La segunda fase es la optimización de  $\mu_k$  dejando  $r_{nk}$  fijo. La función objetivo  $J$  es una función cuadrática de  $\mu_k$ , por tanto, derivando con respecto a  $\mu_k$  e igualando a cero,

$$\frac{\partial J}{\partial \mu_k} = -2 \sum_{n=1}^N \sum_{k=1}^K r_{nk} (x_n - \mu_k) = \sum_{n=1}^N r_{nk} (x_n - \mu_k) = 0$$

Se obtiene el valor optimo  $\mu_k^*$  que minimiza  $J$ , este valor es dado por,

$$\mu_k^* = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}}$$

El denominador de esta expresión es igual al número de puntos correspondientes al grupo  $k$ . Por tanto  $\mu_k^*$  es el centroide del conjunto de puntos  $x_n$  pertenecientes al grupo  $k$  [2].

### 4.3.1. K-Means para el fenotipo Diabetes tipo II

En la sección 4.2.1 se mostro como se puede representar los SNPs informativos en un espacio euclidiano  $d$ -dimensional con  $d \in \{2, 3\}$ , utilizando esta representación para el fenotipo Diabetes tipo II y por medio del procedimiento 7 se calculan los conglomerados de SNPs en  $\mathbb{R}^2$  y  $\mathbb{R}^3$ .

El procedimiento K-Means genera centroides que no necesariamente son SNPs, por tanto, para cada centroide calculado  $\mu_k$  se encuentra el SNP que sea más cercano a este. Esto genera un  $SNP_k$  centroide para el conglomerado  $k$ , este SNP será el representante de la clase  $k$ .

Para efectos prácticos y buscando una mejor visualización de los resultados se corre el procedimiento K-Means con  $N = 100$  y  $k = 5$ , los cálculos y las representaciones se obtuvieron a través del software libre R [18] y la librería `stats`.

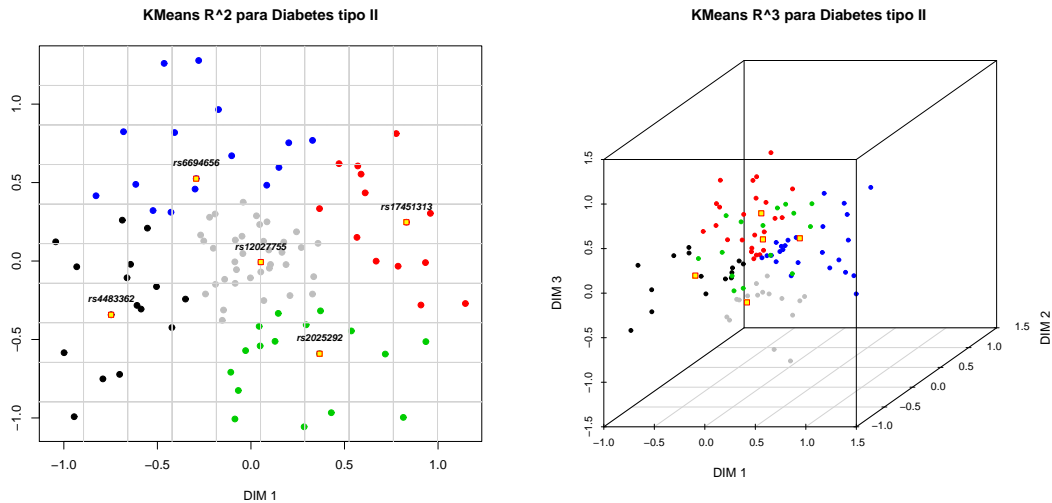
Uno de las limitantes que tienen los estudios de asociación del genoma completo, es el gran volumen de información que se debe administrar para poder encontrar relaciones entre marcadores genéticos dentro del genoma. Este procedimiento permite generar clases de

**Procedimiento 7** Procedimiento de conglomeración K-Means

**Entrada:**  $X = \{x_1, x_2, \dots, x_n\}$  Conjunto de puntos en  $\mathbb{R}^d$   
 $k$  Número de grupos o conglomerados  
 $maxit$  Número máximo de iteraciones

**Salida:**  $C = \{\mu_1, \mu_2, \dots, \mu_k\}$  Número de centroides  
 $L(x) = \{l(x) | x = 1, 2, \dots, n\}$  Etiquetas de grupo para  $X$

- 1: **para todo**  $\mu_i \in C$  **hacer**  
 $\mu_i = x_j \in X$  Se escoge aleatoriamente un conjunto de centroides iniciales
- 2: **fin para**
- 3: **para todo**  $\mu_i \in C$  **hacer**
- 4:   **para todo**  $x_j \in X$  **hacer**  
 $d_j = \|x_j - \mu_i\|$  Distancia entre puntos de  $X$  centroides iniciales
- 5:   **fin para**  
 $l_d(x_i) = \min(d_j)$  ( $j = 1, 2, \dots, n$ ) Distancia mínima con respecto al centroide  
 $l(x_i) = i$  Se asigna etiqueta al punto
- 6: **fin para**  
 $cambio = falso$   
 $iter = 0$
- 7: **repetir**
- 8:   **para todo**  $\mu_i \in C$  **hacer**  
 $ActualizarGrupo(\mu_i)$
- 9:   **fin para**
- 10:   **para todo**  $\mu_i \in C$  **hacer**
- 11:     **para todo**  $x_j \in X$  **hacer**  
 $d_j = \|x_j - \mu_i\|$
- 12:     **fin para**  
 $minDist = \min(d_j)$  ( $j = 1, 2, \dots, n$ )
- 13:     **si**  $minDist \neq l_d(x_i)$  **entonces**
- 14:        $l_d(x_i) = minDist$
- 15:        $cambio = verdadero$
- 16:     **fin si**  
 $l(x_i) = i$
- 17:   **fin para**  
 $iter ++$
- 18: **hasta que**  $cambio = verdadero \wedge iter \leq maxit$



**Figura 4-2:** Agrupación K-Means de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  para Diabetes tipo II.

marcadores disminuyendo en gran medida el número de SNPs que se deben tener en cuenta en un estudio de asociación. En la tabla 4-4 se muestran el conjunto de centroides SNPs asociados al fenotipo Diabetes tipo II. De igual forma en la tabla 4-5 se muestra la configuración de SNPs en  $\mathbb{R}^2$  y  $\mathbb{R}^3$  asociados a los grupos correspondientes a los centroides SNPs. Las configuraciones y salidas del método K-Means para los otros fenotipos se muestran en el apéndice 9.

**Tabla 4-4:** Centroides SNPs para el fenotipo Diabetes tipo II en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs12027755	1	0,05243	-0,00674	rs11809462	1	-0,12829	-0,01856	-0,5405
rs4483362	2	-0,7472	-0,34246	rs17105632	2	-0,49184	-0,28341	-0,1209
rs17451313	3	0,83106	0,2474	rs10874230	3	-0,14133	0,26102	0,33515
rs2025292	4	0,36677	-0,59051	rs7412036	4	0,18038	-0,28963	0,28727
rs6694656	5	-0,2926	0,52507	rs12138259	5	0,31874	0,11763	0,1187

**Tabla 4-5:** Conglomeración Kmeans para el fenotipo Diabetes tipo II en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs278849	1	-0,14054	-0,11657	rs1780049	1	0,05521	-0,53132	-0,27755
rs1166702	1	-0,19152	0,29987	rs17101412	1	-0,07695	0,0793	-0,50233
rs17101412	1	-0,25229	0,12766	rs2025292	1	0,52609	-0,25398	-0,4177
rs11809462	1	-0,26902	0,16584	rs7533564	1	-0,29873	-0,06998	-0,68053
rs12068437	1	-0,22109	0,27896	rs11809462	1	-0,12829	-0,01856	-0,5405
rs273231	1	-0,00292	-0,1072	rs11588580	1	0,41014	-0,17364	-0,6142
rs7553632	1	0,20841	-0,22015	rs4650462	1	0,29582	-0,0072	-1,20201
rs4417015	1	0,16582	0,05548	rs356294	1	0,12455	-0,19325	-0,34818
rs7533564	2	-0,55344	0,20952	rs278849	2	-0,13942	-0,13867	-0,02169
rs11590908	2	-0,66253	-0,10677	rs1539737	2	-0,76376	0,30895	-0,39228
rs1322931	2	-0,68752	0,26043	rs2032056	2	-0,40929	0,21858	-0,31048
rs1608573	2	-0,79145	-0,75107	rs17101082	2	-0,51095	0,39112	-0,44742
rs2043802	2	-0,94489	-0,99205	rs1322931	2	-0,59039	0,08175	-0,48886
rs17467359	2	-0,9311	-0,03665	rs1608573	2	-0,74582	-0,60228	-0,38571
rs4618922	2	-0,60909	-0,28253	rs2043802	2	-0,86341	-0,76354	-0,52347
rs4268311	2	-0,58727	-0,30623	rs4618922	2	-0,62006	-0,1684	0,14334
rs11577132	3	0,61038	0,43448	rs6662245	3	-0,59282	0,99021	0,42239
rs10158555	3	0,90842	-0,28037	rs1166702	3	-0,24399	0,26948	-0,09927
rs3015047	3	1,14614	-0,27079	rs10873963	3	-0,29722	0,45306	-0,21842
rs10493627	3	0,36653	0,33387	rs638335	3	-0,63335	1,05219	0,15398
rs12037276	3	0,47068	0,62051	rs6694656	3	-0,40912	0,42913	0,24877
rs17105038	3	0,95897	0,30388	rs11590908	3	-0,44647	0,01072	0,55563
rs17415150	3	0,9327	-0,00999	rs12068437	3	-0,2055	0,24294	-0,16644
rs713332	3	0,66892	-0,00079	rs273231	3	-0,02892	-0,11031	0,10986
rs1780049	4	0,04904	-0,54094	rs17101646	4	0,36821	-0,62743	0,59399
rs2025292	4	0,36677	-0,59051	rs1929984	4	0,04179	-0,59045	0,0112
rs17101646	4	0,71792	-0,59274	rs1413357	4	-0,12191	-0,89493	0,34168
rs1929984	4	-0,02856	-0,57088	rs7553632	4	-0,04819	-0,26712	0,47477
rs1413357	4	-0,08581	-1,0067	rs12121256	4	0,2734	-0,32134	0,12394
rs1487546	4	0,37164	-0,31748	rs17105038	4	0,64527	-0,25364	0,67204
rs11588580	4	0,53653	-0,4455	rs17098717	4	0,22725	0,03138	0,53977
rs4650462	4	0,81438	-0,99645	rs7412036	4	0,18038	-0,28963	0,28727
rs6662245	5	-0,27936	1,27797	rs11577132	5	0,66714	0,35534	-0,00703
rs1539737	5	-0,82816	0,41588	rs4417015	5	0,20933	0,05321	0,06516
rs2032056	5	-0,42753	0,31097	rs1487546	5	0,44328	-0,25899	-0,13461
rs17101082	5	-0,61559	0,48859	rs10158555	5	0,90913	-0,12912	-0,15292
rs10873963	5	-0,29875	0,45856	rs12027492	5	0,32747	0,07166	-0,13351
rs638335	5	-0,46399	1,25959	rs3015047	5	1,01175	-0,12365	-0,39804
rs6694656	5	-0,2926	0,52507	rs6657674	5	0,03637	0,56233	-0,06973
rs6657674	5	0,08441	0,48321	rs10493627	5	0,38875	0,39409	-0,16245

# 5 Construcción de Redes de SNPs entre Fenotipos

El objetivo de los estudios de asociación entre enfermedades es encontrar factores genéticos correlacionados con enfermedades complejas. En estos estudios, se realiza un proceso de muestreo del DNA de dos poblaciones (*casos - controles*). Las asociaciones entre las estructuras de los haplotipos de las dos poblaciones son relevantes para realizar pruebas estadísticas. Estas asociaciones sirven como evidencia para medir la correlación de las regiones genómicas estudiadas para la enfermedad.

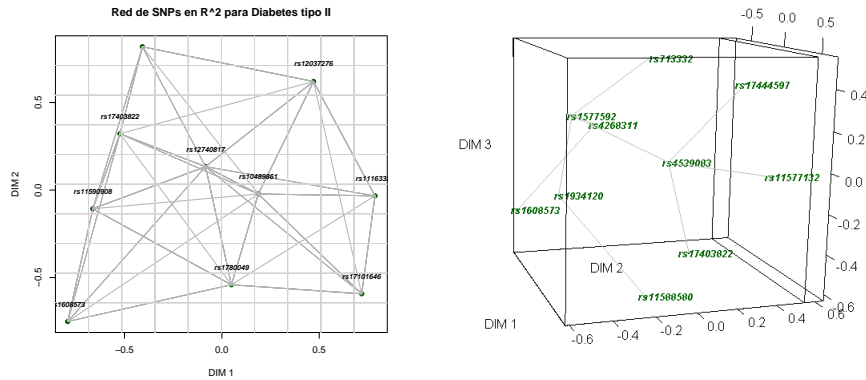
La significancia estadística del estudio está directamente asociada al número de SNPs seleccionados. Por tanto, una selección de un número pequeño de SNPs implica un ahorro de recursos. Esto se hace, como se ha mencionado en capítulos anteriores, eligiendo un subconjunto adecuado de SNPs, conocido como SNPs etiqueta. Este proceso permite al investigador centrarse solo en las etiquetas para encontrar asociaciones entre enfermedades, lo cual, representa un ahorro considerable de recursos. Así mismo, la potencia de las pruebas estadísticas que se realizan aumenta directamente al número de SNPs etiqueta seleccionados [8].

## 5.1. Construcción de redes de SNPs para un mismo fenotipo

El análisis de escalamiento multidimensional permite representar la distribución de los elementos del espacio de SNPs en un espacio euclidiano, en nuestro caso  $\mathbb{R}^2$  y  $\mathbb{R}^3$ . Esto facilita la visualización de los SNPs presentes en los haplotipos. El principal interés de representar los SNPs en un espacio vectorial, es poder identificar conglomerados de marcadores genéticos y seleccionar los centroides de estos clusters, como los SNPs etiquetas de interés (fig 5-1).

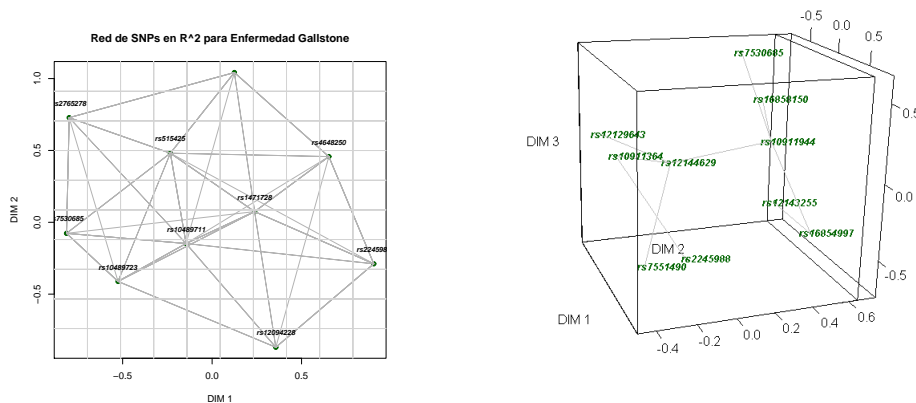
La metodología de conglomeración por  $K$ -Means es una metodología no paramétrica, iterativa, estable y de rápida convergencia. Estos tipos de métodos numéricos tienen la ventaja de evitar la estimación de parámetros para ingresar al modelo de conglomeración, pero a su vez, también es una limitante, ya que para cada iteración el mínimo de la función objetivo en el proceso de  $K$ -Means se puede alcanzar al estimar diferentes configuraciones y escogencias de centroides. Por tanto dependiendo del número de iteraciones los SNPs etiquetas pueden cambiar dentro del cluster (fig 5-2).

Esta pequeña y única limitante no es un impedimento para el uso de esta metodología ya



**Figura 5-1:** Red de SNPs en  $\mathbb{R}^2$  y  $\mathbb{R}^3$  asociados al fenotipo Diabetes tipo II.

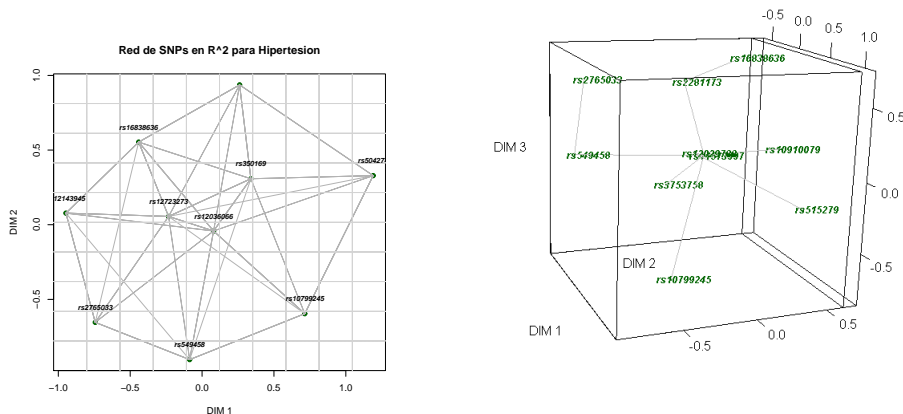
que la inestabilidad en la escogencia de los SNPs centroides o etiquetas es debido al conjunto pequeño de haplotipos considerados en el estudio. Dado que el volumen de haplotipos almacenados en bases de datos públicas es gigante la potencia de esta metodología aumenta al considerar un conjunto grande de conglomerados. Sin embargo, el tiempo de ejecución y número de iteraciones aumenta de forma proporcional (fig 5-3).



**Figura 5-2:** Red de SNPs en  $\mathbb{R}^2$  y  $\mathbb{R}^3$  asociados al fenotipo Enfermedad Gallstone.

Una vez encontrado el subconjunto de SNPs etiqueta se pueden establecer relaciones entre los conglomerados solo examinando la correlación existente entre marcadores o etiquetas SNPs, como se observa en las figuras 5-1 a 5-3. La convergencia del método de escalamiento multidimensional no métrico, depende de la minimización de la estadística *Stress*. Para cada uno de los fenotipos se observa que las mejores representaciones se logran considerando el espacio  $\mathbb{R}^3$ , esto también corrobora la disposición de los SNPs centroides que facilitan la

formación de la red entre conglomerados.



**Figura 5-3:** Red de SNPs en  $\mathbb{R}^2$  y  $\mathbb{R}^3$  asociados al fenotipo Hipertensión.

Definido el conjunto de SNPs etiqueta para cada uno de los fenotipos, DII: Diabetes tipo II, GLS: Enfermedad de Gallstone e HIP: Hipertensión, se puede establecer una red entre los representantes de cada conglomerado. Esta red global entre fenotipos, permite encontrar asociaciones entre enfermedades con la identificación de marcadores y la conexión existente en el espacio de representación (fig 5-4). La distribución espacial de los SNPs garantiza encontrar caminos óptimos entre marcadores y le facilita al investigador la visualización de estas asociaciones ahorrando recursos en la búsqueda de sectores del genoma correspondientes a los marcadores SNPs. Las coordenadas de los SNPs etiquetas asociados a cada fenotipo se muestran en la tabla 9-5.

## 5.2. Relación física de SNPs a lo largo del cromosoma

Al seleccionar un subconjunto óptimo de SNPs de un grupo de casos correlacionados con un grupo de control, se puede encontrar sectores físicos en el genoma que estén asociados a una enfermedad particular.

Alelos de SNPs en estrecha proximidad física a menudo están correlacionados y la variación de la secuencia de los alelos en lugares contiguos a lo largo de la región cromosómica marcada por el SNP tiene diversidad limitada [8].

El algoritmo de  $K$ -Means permite separar grupos de SNPs en el espacio de representación. Cada clase de SNPs queda determinada por el SNP etiqueta o centroide del grupo. Estos a su vez, tienen asociado una posición a lo largo del genoma, lo que permite encontrar cúmulos de SNPs asociando sus sectores físicos, los cuales son los relevantes para determinar el desenlace



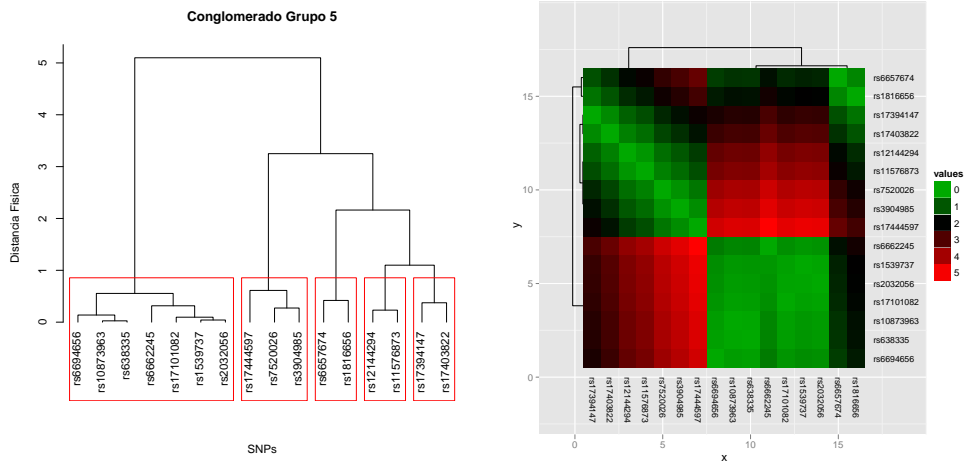


Figura 5-5: Conglomerado para el grupo 5 asociado al fenotipo Diabetes tipo II.

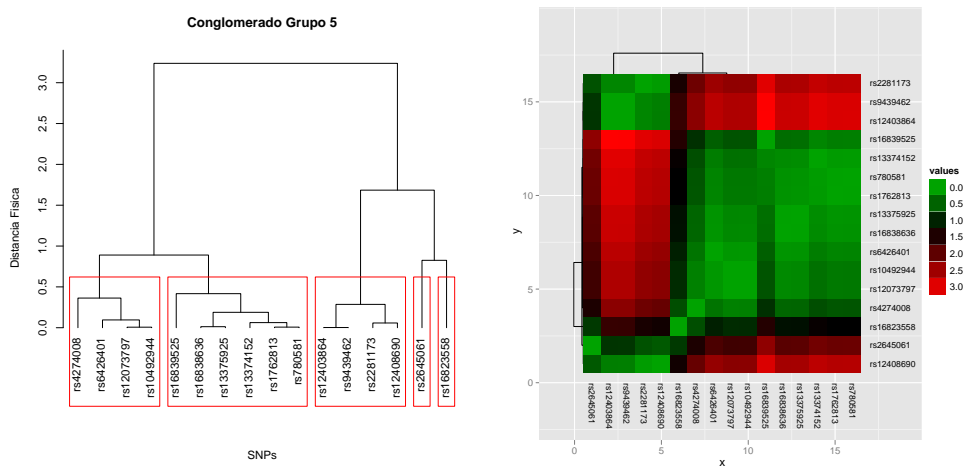


Figura 5-6: Conglomerado para el grupo 5 asociado al fenotipo Hipertensión.

a localizar características genéticas entre enfermedades, con la ayuda de la visualización de la red global entre marcadores y la selección de SNPs que tengan una alta correlación con su proximidad física.

## 6 Discusión General de los Resultados Obtenidos

La importancia de desarrollar metodologías para los estudios de asociación del genoma completo (GWAS), radica en presentar diferentes soluciones para el tratamiento de los datos de marcadores genéticos SNPs. Una de las limitantes de estos estudios es el gran volumen de información que se debe manipular, permitiendo la obtención de falsos positivos que obstruye una posible explicación para el desenlace y tratamiento de una enfermedad. Halperin et al. en [8] desarrollan un algoritmo para la predicción de SNPs representantes de un bloque de haplotipos, buscando reducir el número de SNPs para el estudio de asociación, con el agravante de que el SNP etiqueta puede no pertenecer al haplotipo. Zelikovsky et al. en [13] y [12] desarrollan algoritmos para definir un espacio vectorial de SNPs, suponiendo que cada SNP es elemento de  $\mathbb{Z}_3^n$ , para luego encontrar los marcadores etiquetas midiendo las distancias entre elementos del espacio con la métrica usual en  $\mathbb{Z}_3^n$  como subespacio de  $\mathbb{R}^n$ , este algoritmo tiene el inconveniente que el espacio de SNPs no cumple con algunos axiomas de espacio vectorial. En este trabajo se presenta una metodología que sigue lineamientos estadísticos, realizando sucesivos filtros para reducir la información irrelevante, definiendo una medida de correlación entre SNPs de dos poblaciones distintas, en este caso (Enfermos: Casos y Sanos: Controles), de tal forma que al encontrar los marcadores genéticos altamente correlacionados, se reduce de forma optima el numero de marcadores relevantes para un GWAS.

Tradicionalmente en un GWAS, para encontrar asociaciones entre las muestras casos-control se seleccionan los mismos SNPs presentes en la población sana y enferma. Se realiza un conteo de alelos para cada SNP y se calcula una estadística la cual se supone que se distribuye  $\chi^2$ .

**Tabla 6-1:** Conteo de alelos presentes en SNPs.

Población	a/a	A/a	A/A	Total
Caso	$n_{11}$	$n_{12}$	$n_{13}$	$n_{1\bullet}$
Control	$n_{21}$	$n_{22}$	$n_{23}$	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	$n$

Sobre esta tabla de contingencia se calcula la estadística.

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(n_{ij} - E[n_{ij}])^2}{E[n_{ij}]}$$

Donde  $E[n_{ij}] = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ , con  $X^2 \sim \chi_{(2)}^2$ , luego se le asocia el  $p$ -valor correspondiente. La representación grafica de estos datos quedan determinados por su posición física y su correspondiente  $p$ -valor para cada SNP. Esta representación es limitada ya que depende del supuesto de distribución  $\chi^2$  y solo da información sobre la frecuencia de los alelos. Zelikovsky et al. en [13] y [12] dan un perspectiva multivariada para la representación de SNPs en un espacio euclidiano, suponiendo que el espacio de SNPs es subespacio vectorial de  $\mathbb{R}^n$ . Dentro de los algoritmos desarrollados en este trabajo se muestra que el espacio de SNPs posee estructura de espacio métrico, por tanto surge la posibilidad de obtener una representación de los elementos del espacio de SNPs en  $\mathbb{R}^n$  utilizando escalamiento multidimensional. Esta técnica multivariada garantiza la representación y visualización de conglomerados de SNPs, permitiendo dar un marco multivariado al análisis de asociaciones e identificación de sectores del genoma relacionados con el desenlace de una enfermedad.

En un GWAS promedio, las asociaciones se establecen entre los mismos marcadores, eliminando la posibilidad de correlación entre SNPs pertenecientes a los mismos o diferentes fenotipos. Al considerar el espacio de SNPs como un espacio métrico, se puede definir una medida de correlación a partir de las distancias entre elementos de este espacio. Por tanto, al considerar uno o varios fenotipos, se pueden establecer relaciones entre SNPs de diferentes partes físicas a lo largo del genoma. Para encontrar estas asociaciones, se encuentran conglomerados de SNPs en el espacio euclidiano de representación, para luego determinar si estos mismos conservan esta relación de cercanía en los sectores físicos del genoma asociados a sus posiciones.

Desde el año 2010 el instituto de genética de la Universidad Nacional de Colombia en colaboración con investigadores de la Pontificia Universidad Católica de Chile, vienen desarrollando un proyecto de investigación para clasificar y reconstruir redes de SNPs por factores de riesgo para el diagnostico de enfermedades. La metodología desarrollada en este trabajo permite dar un aporte desde el punto de vista estadístico, para encontrar marcadores genéticos relevantes en un estudio de asociación.

# 7 Conclusiones y recomendaciones

## 7.1. Conclusiones

Las metodologías desarrolladas para los estudios de asociación del genoma completo (GWAS) seguirán actualizándose, aplicando nuevas técnicas estadísticas para encontrar asociaciones entre marcadores SNPs. Inicialmente se consideraban solo las frecuencias de alelos presentes en a lo largo del genoma, ahora con esta metodología se pueden representar los marcadores genéticos en un espacio vectorial euclidiano, garantizando el ahorro de recursos en la búsqueda de conglomerados de SNPs.

Las asociaciones entre poblaciones casos-control pueden ser considerados entre marcadores de distintos fenotipos, estimando así, otros sectores del genoma que pueden ser relevantes en un GWAS, cumpliendo con la afirmación empírica de que la alta correlación entre marcadores es proporcional a la cercanía de los mismos a lo largo del genoma.

El proceso de identificación de SNPs etiqueta por medio de técnicas de aprendizaje no supervisado resulta ser mas optimo, ya que al no existir estimación de parámetros para la separación de grupos en el espacio de representación, el proceso iterativo escoge las clases y sus respectivos centroides o representantes de clase. Ahorrando así, supuestos sobre el espacio de SNPs que no siempre se cumplen.

Generalmente los estudios de asociación consideran los mismos marcadores SNPs presentes en las muestras casos-control, asumiendo que las características genéticas asociadas a una enfermedad conservan los mismos sectores del genoma. Contrariamente a lo esperado, se pueden obtener distintos sectores genéticos que pueden influir en el desenlace de una enfermedad, ubicando la posición física de los SNPs pertenecientes a los conglomerados obtenidos en el proceso de agrupación y separación sobre el espacio de representación.

## 7.2. Recomendaciones

Las técnicas multivariadas desarrolladas en este trabajo para la representación de SNPs en espacios euclidianos, pueden ser utilizadas para encontrar conglomerados de SNPs haciendo uso de resultados obtenidos en minería de datos. Técnicas como *Support Vector Machines* (SVM) pueden ser utilizadas para encontrar conglomerados de SNPs , garantizando la selección de SNPs etiqueta de forma optima y estable, eliminado la posibilidad de alteraciones

de los elementos de clase debido al número de iteraciones o tamaño de la muestra. Las asociaciones entre fenotipos de SNPs fueron obtenidas a partir de una medida de correlación definida en el espacio métrico de SNPs. También se pueden considerar las discrepancias, seleccionando aquellos marcadores entre las poblaciones casos-control que tengan un nivel de correlación bajo.

# 8 EMD-NM para los fenotipos Enfermedad de Gallstone e Hipertensión

**Tabla 8-1:** EMD-NM para el fenotipo Enfermedad Gallstone en  $R^2$  y  $R^3$ .

Num Dim	Stress NM	Num Iter
2	0,05248	256
3	0,02525	631

**Tabla 8-2:** Configuración en  $R^2$  y  $R^3$  para el fenotipo Enfermedad Gallstone.

SNP	D1	D2	SNP	D1	D2	D3
rs16854997	0,0884	0,83682	rs16854997	0,05699	0,63281	-0,55058
rs12137562	0,38583	-0,16591	rs12137562	0,22746	-0,03009	-0,44432
rs7516257	0,61909	-0,74718	rs7516257	0,40352	-0,81169	-0,18788
rs12036985	0,93238	0,2959	rs12036985	0,85825	0,04415	0,37181
rs7556324	0,49578	-0,54713	rs7556324	0,39789	-0,60664	-0,10858
rs12145255	-0,92976	0,07092	rs12145255	-0,43402	0,11634	0,8198
rs12760195	0,1624	0,16019	rs12760195	0,15596	0,18277	-0,11314
rs34750348	-0,1697	1,31872	rs34750348	-0,69809	0,41531	-0,9393
rs12040223	-0,28451	-0,17452	rs12040223	-0,30976	-0,19521	0,02231
rs4652537	-0,12743	0,18517	rs4652537	-0,05143	0,2839	0,18406
rs4651077	-0,14907	0,06611	rs4651077	-0,15118	0,04945	0,11259
rs12125196	-0,16917	-0,52273	rs12125196	-0,12652	-0,49558	-0,41824
rs12060051	-0,02589	-0,79256	rs12060051	0,04814	-0,50833	0,58272
rs12238995	-0,95489	-0,69646	rs12238995	-0,92146	-0,54253	-0,30622
rs10494537	0,12834	-0,13895	rs10494537	0,0499	-0,03803	0,2145
rs12565063	-0,47962	0,52839	rs12565063	-0,5122	0,12716	-0,56218
rs10910900	0,62116	-0,09515	rs10910900	0,65095	0,09898	-0,06577
rs6703292	0,25263	0,14475	rs6703292	0,27681	0,17512	-0,05201
rs16857194	-0,109	-0,00921	rs16857194	-0,11545	0,00413	0,02613
rs583757	-0,06686	-0,03276	rs583757	-0,0582	-0,02494	-0,05715

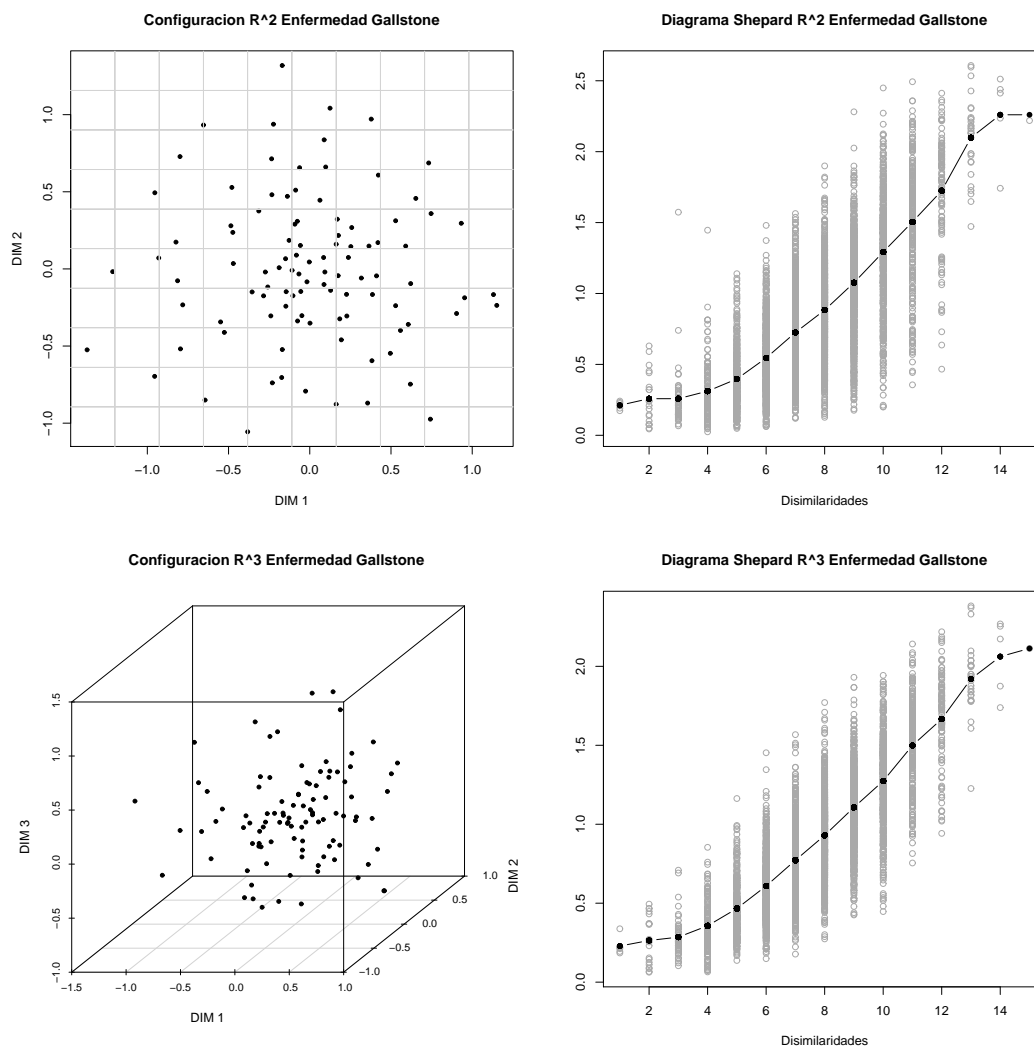


Figura 8-1: Configuración de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  y Diagramas de Shepard.

**Tabla 8-3:** Stress por punto en  $R^2$  y  $R^3$  para el fenotipo Enfermedad Gallstone.

SNP	StPP $R^2$	SNP	StPP $R^3$
rs12040223	0,0084	rs12040223	0,00228
rs11800990	0,01051	rs16823145	0,00399
rs16823145	0,01124	rs11800990	0,00521
rs12760195	0,01228	rs17481405	0,00685
rs10489711	0,01329	rs10489711	0,00711
rs6703292	0,01513	rs4651077	0,00772
rs10911944	0,01528	rs11590854	0,00779
rs11590854	0,01592	rs10494547	0,00815
rs10494547	0,01674	rs12745536	0,00836
rs4651077	0,01732	rs12760195	0,00887
rs583757	0,01846	rs11810323	0,00914
rs17838217	0,01893	rs10911944	0,00996
rs12132118	0,01925	rs10458347	0,01005
rs3845440	0,01993	rs10911364	0,01015
rs16860537	0,02027	rs17838217	0,01035
rs10458347	0,02104	rs16860537	0,01041
rs16858267	0,02166	rs234657	0,01109
rs4652537	0,02169	rs4652537	0,01119
rs17481405	0,02245	rs10489406	0,01124
rs12745536	0,02251	rs6703292	0,01129

**Tabla 8-4:** EMD-NM para el fenotipo Hipertensión en  $R^2$  y  $R^3$ .

Num Dim	Stress NM	Num Iter
2	0,04611	275
3	0,0209	1000

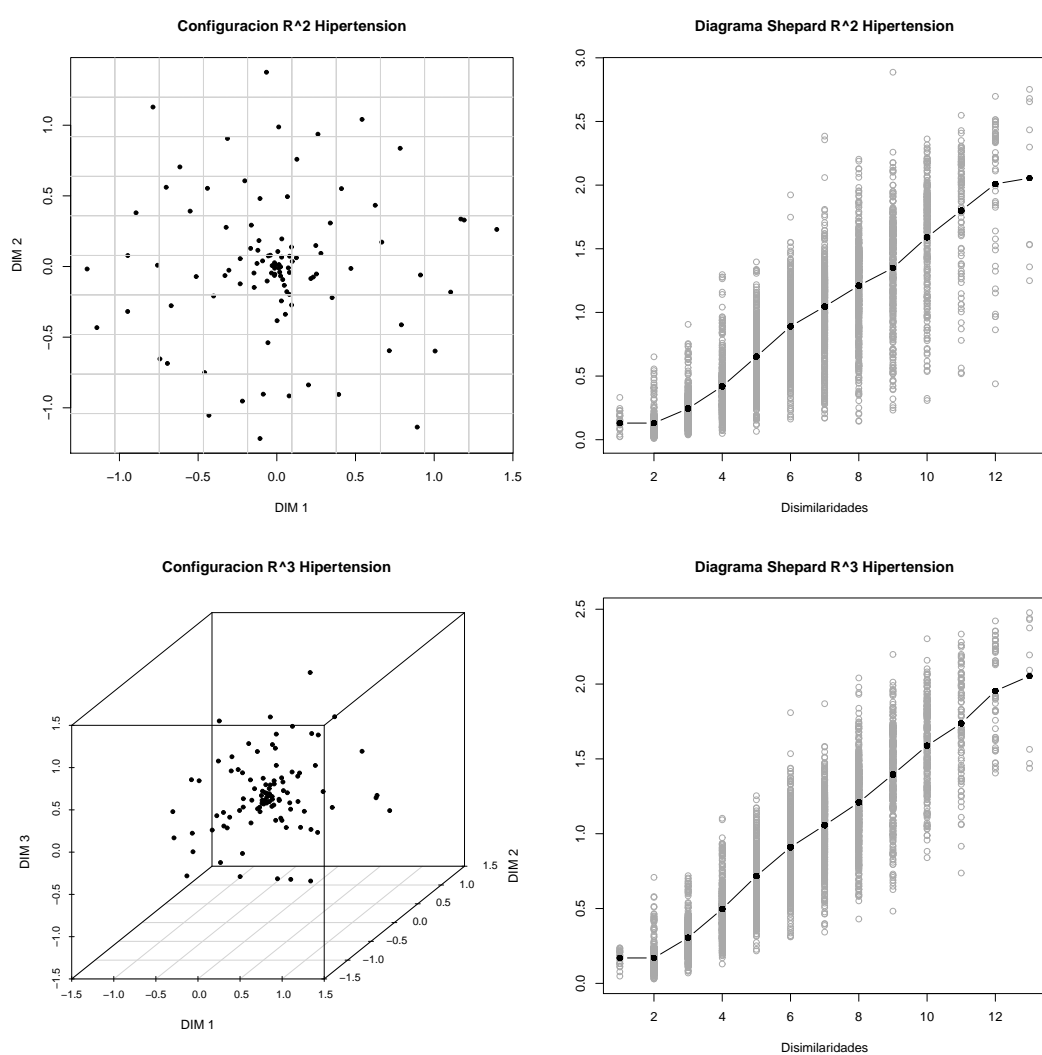


Figura 8-2: Configuración de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  y Diagramas de Shepard.

**Tabla 8-5:** Configuración en  $R^2$  y  $R^3$  para el fenotipo Hipertensión.

SNP	D1	D2	SNP	D1	D2	D3
rs2977656	0,03747	-0,0923	rs2977656	0,03768	-0,10398	-0,02614
rs2519068	0,25157	-0,05294	rs2519068	0,18359	-0,03594	-0,27478
rs4970403	0,27997	0,09324	rs4970403	0,25431	-0,32875	0,28603
rs11260595	0,05404	-0,33813	rs11260595	0,01358	-0,39093	-0,14751
rs1320571	0,89179	-1,13644	rs1320571	0,85352	-0,90277	0,6325
rs3813200	0,62509	0,43343	rs3813200	0,47568	0,51498	-0,3665
rs12073590	-0,69527	-0,68582	rs12073590	-0,54132	-0,66548	0,48578
rs2765033	-0,743	-0,65387	rs2765033	-0,40274	-0,75116	0,51138
rs9439432	-0,04118	0,0799	rs9439432	-0,01089	0,06473	0,0998
rs12403864	-0,31361	0,90622	rs12403864	-0,40005	0,81238	0,24473
rs9439462	-0,06594	1,37499	rs9439462	-0,18716	1,2685	0,17267
rs880050	0,01827	-0,03993	rs880050	0,02891	-0,04906	-0,04886
rs4970458	0,12421	0,06213	rs4970458	0,09214	0,16136	-0,01079
rs909823	0,03152	0,19542	rs909823	-0,0389	0,19744	0,08979
rs2281173	-0,1071	0,48111	rs2281173	-0,43267	0,00442	0,45967
rs11582768	0,03062	0,06628	rs11582768	0,19883	-0,04176	0,18181
rs12408690	-0,16238	0,29278	rs12408690	-0,34823	-0,01444	0,31667
rs9786963	-0,12062	0,11494	rs9786963	-0,20656	-0,00665	0,19104
rs6664578	-0,05394	0,07486	rs6664578	-0,06702	0,0709	0,09904
rs2490561	0,04732	-0,13312	rs2490561	-0,03253	-0,16675	-0,07988

**Tabla 8-6:** Stress por punto en  $R^2$  y  $R^3$  para el fenotipo Hipertensión.

SNP	StPP $R^2$	SNP	StPP $R^3$
rs16840868	0,0024	rs16840868	0,00095
rs4648646	0,00427	rs4648646	0,00215
rs4654535	0,00524	rs12073856	0,00228
rs880050	0,00528	rs880050	0,00248
rs11573997	0,00533	rs11808208	0,00275
rs12073856	0,00541	rs11573997	0,00284
rs2817181	0,00578	rs12408633	0,00321
rs11808208	0,00624	rs9786963	0,00321
rs12026862	0,00631	rs2817181	0,0034
rs2651906	0,00652	rs12046130	0,00346
rs16823103	0,00692	rs12026862	0,0035
rs11588666	0,00707	rs4654535	0,00361
rs12046130	0,00707	rs2651906	0,00362
rs7550609	0,00724	rs7550609	0,00405
rs12038607	0,00728	rs12036066	0,00408
rs747778	0,00734	rs974635	0,00415
rs10909904	0,00748	rs12038607	0,0042
rs16824328	0,0076	rs16824328	0,00427
rs974635	0,00765	rs2977656	0,00473
rs12036066	0,00809	rs10909904	0,00499

# 9 K-Means para los fenotipos Enfermedad de Gallstone e Hipertensión

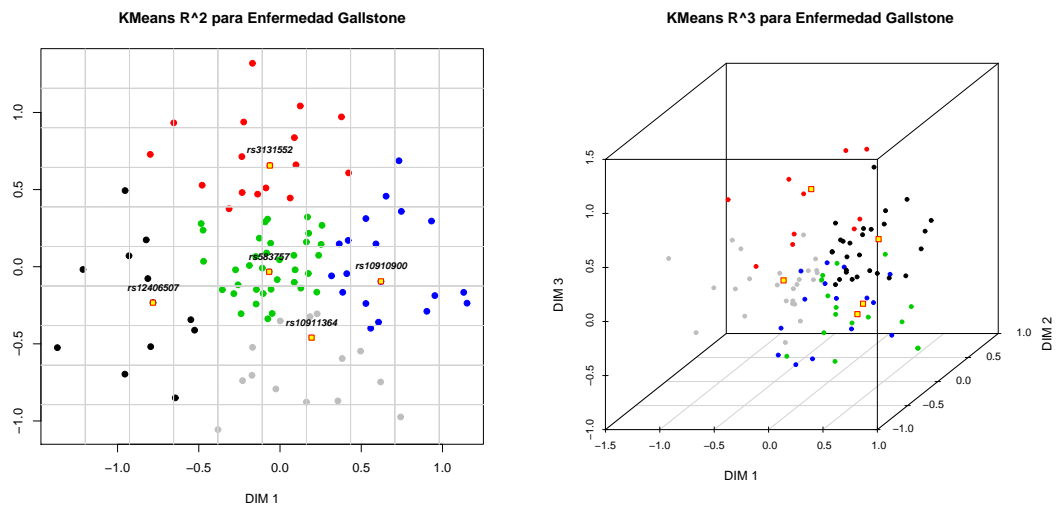
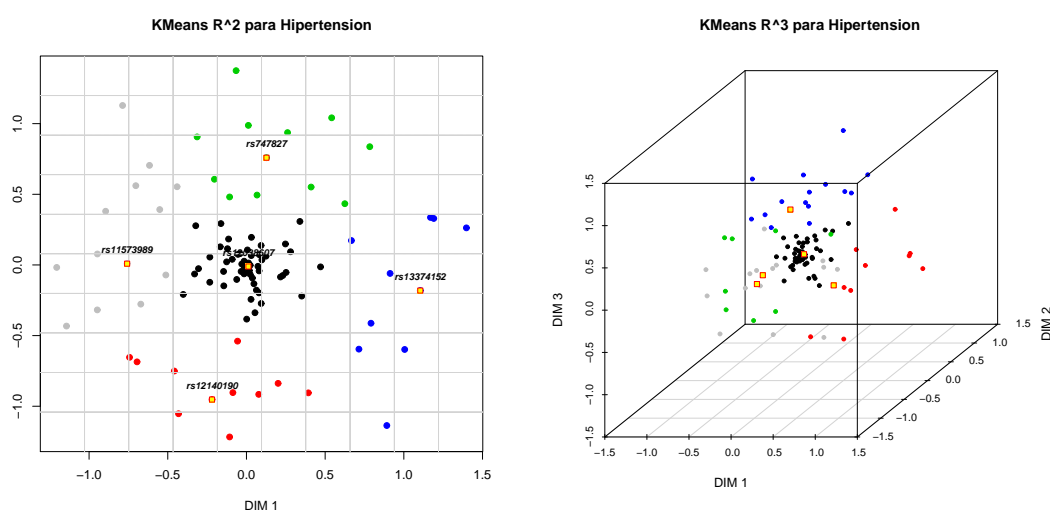


Figura 9-1: Agrupación K-Means de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  para Enfermedad de Gallstone.

Tabla 9-1: Centroides SNPs para el fenotipo Enfermedad Gallstone en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs10911364	1	0,19405	-0,45974	rs12040223	1	-0,30976	-0,19521	0,02231
rs12406507	2	-0,78205	-0,23307	rs10489402	2	0,27414	0,32576	0,17228
rs3131552	3	-0,0627	0,65609	rs7530685	3	-0,3282	0,29667	0,6483
rs583757	4	-0,06686	-0,03276	rs17494103	4	0,50736	-0,44261	-0,17988
rs10910900	5	0,62116	-0,09515	rs3818805	5	0,10151	0,3801	-0,44846



**Figura 9-2:** Agrupación K-Means de SNPs en  $\mathbb{R}^2$ ,  $\mathbb{R}^3$  para Hipertensión.

**Tabla 9-2:** Centroides SNPs para el fenotipo Hipertension en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs11573989	1	-0,75826	0,009	rs12138240	1	-0,4804	0,04307	-0,2728
rs12038607	2	0,01261	-0,01097	rs11573997	2	0,02925	0,00764	-0,00849
rs12140190	3	-0,21886	-0,95249	rs2071992	3	0,38516	0,00034	-0,37358
rs747827	4	0,12721	0,75897	rs549458	4	-0,02099	-0,90937	0,0461
rs13374152	5	1,1046	-0,18111	rs4274008	5	-0,22824	0,18053	0,44338

**Tabla 9-3:** Conglomeración Kmeans para el fenotipo Enfermedad Gallstone en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs7516257	1	0,61909	-0,74718	rs12040223	1	-0,30976	-0,19521	0,02231
rs7556324	1	0,49578	-0,54713	rs4651077	1	-0,15118	0,04945	0,11259
rs12125196	1	-0,16917	-0,52273	rs12125196	1	-0,12652	-0,49558	-0,41824
rs12060051	1	-0,02589	-0,79256	rs12060051	1	0,04814	-0,50833	0,58272
rs7551490	1	-0,23066	-0,73863	rs12238995	1	-0,92146	-0,54253	-0,30622
rs6688919	1	0,16234	-0,87736	rs16857194	1	-0,11545	0,00413	0,02613
rs12125461	1	0,38252	-0,59521	rs583757	1	-0,0582	-0,02494	-0,05715
rs12094228	1	0,35613	-0,86976	rs12752411	1	-0,1862	-0,26865	-0,15951
rs12145255	2	-0,92976	0,07092	rs12036985	2	0,85825	0,04415	0,37181
rs12238995	2	-0,95489	-0,69646	rs12760195	2	0,15596	0,18277	-0,11314
rs12089943	2	-0,95382	0,49358	rs4652537	2	-0,05143	0,2839	0,18406
rs10494534	2	-0,64461	-0,8502	rs10494537	2	0,0499	-0,03803	0,2145
rs12129643	2	-0,79616	-0,51835	rs10910900	2	0,65095	0,09898	-0,06577
rs7541043	2	-1,37085	-0,52467	rs6703292	2	0,27681	0,17512	-0,05201
rs12406507	2	-0,78205	-0,23307	rs596424	2	0,43225	0,21668	-0,13816
rs12078468	2	-1,21313	-0,01787	rs3845440	2	0,10167	0,19934	-0,14293
rs16854997	3	0,0884	0,83682	rs12145255	3	-0,43402	0,11634	0,8198
rs34750348	3	-0,1697	1,31872	rs1779799	3	-0,48318	0,29158	0,23514
rs12565063	3	-0,47962	0,52839	rs12089943	3	-0,63471	0,7222	0,41567
rs2765278	3	-0,79892	0,72821	rs2494457	3	-0,47926	0,26086	0,15278
rs6704320	3	0,37816	0,97131	rs2765278	3	0,03595	0,21303	1,04261
rs3818805	3	0,06247	0,44539	rs12078468	3	-0,90357	-0,04373	0,70196
rs2050524	3	0,09802	0,66061	rs16860731	3	-0,07946	0,55932	0,16359
rs16860731	3	-0,13792	0,47056	rs13376315	3	-0,81098	0,25191	-0,04668
rs12760195	4	0,1624	0,16019	rs12137562	4	0,22746	-0,03009	-0,44432
rs12040223	4	-0,28451	-0,17452	rs7516257	4	0,40352	-0,81169	-0,18788
rs4652537	4	-0,12743	0,18517	rs7556324	4	0,39789	-0,60664	-0,10858
rs4651077	4	-0,14907	0,06611	rs681590	4	0,19465	-0,25192	-0,70092
rs10494537	4	0,12834	-0,13895	rs17494103	4	0,50736	-0,44261	-0,17988
rs6703292	4	0,25263	0,14475	rs3021389	4	0,88011	-0,22343	-0,20731
rs16857194	4	-0,109	-0,00921	rs1689782	4	0,89776	-0,15025	-0,62454
rs583757	4	-0,06686	-0,03276	rs7543416	4	0,8522	-0,05919	-0,66482
rs12137562	5	0,38583	-0,16591	rs16854997	5	0,05699	0,63281	-0,55058
rs12036985	5	0,93238	0,2959	rs34750348	5	-0,69809	0,41531	-0,9393
rs10910900	5	0,62116	-0,09515	rs12565063	5	-0,5122	0,12716	-0,56218
rs596424	5	0,4195	0,17056	rs16858267	5	-0,10558	0,12311	-0,14852
rs681590	5	0,60622	-0,35961	rs7551490	5	-0,15787	-0,26565	-0,7256
rs17494103	5	0,55812	-0,39894	rs4652698	5	0,10892	0,17491	-0,59071
rs16858150	5	0,52846	0,31219	rs6704320	5	0,26791	0,55589	-0,8184
rs3021389	5	0,95307	-0,18717	rs3818805	5	0,10151	0,3801	-0,44846

**Tabla 9-4:** Conglomeración Kmeans para el fenotipo Hipertension en  $R^2$  y  $R^3$ .

SNP	Grupo	D1	D2	SNP	Grupo	D1	D2	D3
rs11573989	1	-0,75826	0,009	rs11573989	1	-0,66492	0,24195	-0,30476
rs3753758	1	-0,67145	-0,27739	rs6677992	1	-0,30844	0,02195	-0,14391
rs16823558	1	-0,55101	0,39198	rs3753758	1	-0,64117	-0,04222	-0,38703
rs12757243	1	-1,14357	-0,43249	rs16823548	1	-0,33703	-0,00416	-0,17221
rs10492944	1	-0,70311	0,56089	rs12058584	1	-0,38419	-0,1848	-0,29803
rs12138240	1	-0,51241	-0,07109	rs2483239	1	-0,21524	-0,21181	-0,86196
rs6426401	1	-0,61611	0,70479	rs3765762	1	-0,01275	0,79628	-0,53762
rs12143945	1	-0,94711	0,07765	rs747827	1	-0,08359	0,63134	-0,43957
rs2977656	2	0,03747	-0,0923	rs2977656	2	0,03768	-0,10398	-0,02614
rs2519068	2	0,25157	-0,05294	rs2519068	2	0,18359	-0,03594	-0,27478
rs4970403	2	0,27997	0,09324	rs4970403	2	0,25431	-0,32875	0,28603
rs11260595	2	0,05404	-0,33813	rs11260595	2	0,01358	-0,39093	-0,14751
rs9439432	2	-0,04118	0,0799	rs9439432	2	-0,01089	0,06473	0,0998
rs880050	2	0,01827	-0,03993	rs880050	2	0,02891	-0,04906	-0,04886
rs4970458	2	0,12421	0,06213	rs4970458	2	0,09214	0,16136	-0,01079
rs909823	2	0,03152	0,19542	rs909823	2	-0,0389	0,19744	0,08979
rs12073590	3	-0,69527	-0,68582	rs3813200	3	0,47568	0,51498	-0,3665
rs2765033	3	-0,743	-0,65387	rs4418531	3	0,67076	-0,2977	-0,87608
rs12140190	3	-0,21886	-0,95249	rs905135	3	0,60865	0,9127	0,12088
rs2483239	3	-0,45852	-0,75079	rs4648392	3	0,7663	-0,4634	-0,19222
rs16823922	3	0,078	-0,91588	rs12029788	3	0,85566	-0,36436	0,21175
rs4292923	3	0,39314	-0,90518	rs10799245	3	0,27608	-0,29682	-0,84937
rs16838560	3	-0,10703	-1,21736	rs241225	3	1,30966	0,24249	-0,2831
rs1762813	3	-0,43165	-1,05298	rs515279	3	1,1202	0,29145	-0,15326
rs3813200	4	0,62509	0,43343	rs1320571	4	0,85352	-0,90277	0,6325
rs12403864	4	-0,31361	0,90622	rs12073590	4	-0,54132	-0,66548	0,48578
rs9439462	4	-0,06594	1,37499	rs2765033	4	-0,40274	-0,75116	0,51138
rs2281173	4	-0,1071	0,48111	rs12140190	4	-0,46172	-0,7765	-0,31712
rs2645061	4	0,41067	0,55126	rs16823922	4	-0,09774	-0,84278	-0,41502
rs10910079	4	0,06711	0,49446	rs4292923	4	0,20982	-0,92608	-0,27078
rs905135	4	0,78346	0,83675	rs16838560	4	-0,2668	-1,14196	0,06476
rs3765762	4	0,26115	0,93659	rs16839003	4	0,09197	-0,71354	0,58919
rs1320571	5	0,89179	-1,13644	rs12403864	5	-0,40005	0,81238	0,24473
rs4418531	5	1,00543	-0,59816	rs9439462	5	-0,18716	1,2685	0,17267
rs4648392	5	0,79127	-0,41251	rs2281173	5	-0,43267	0,00442	0,45967
rs12029788	5	0,91269	-0,06011	rs12408690	5	-0,34823	-0,01444	0,31667
rs10799245	5	0,71441	-0,59589	rs2645061	5	0,33354	0,47099	0,50997
rs13374152	5	1,1046	-0,18111	rs16823558	5	-0,6176	0,04717	0,38998
rs241225	5	1,39749	0,26243	rs4274008	5	-0,22824	0,18053	0,44338
rs515279	5	1,16935	0,33636	rs12073797	5	-0,00335	0,18021	0,27945

**Tabla 9-5:** Configuración de Red de SNPs en  $R^3$  para los distintos fenotipos.

SNP	Fenotipo	D1	D2	D3
rs17444597	DII	-0,24432	0,52922	0,33919
rs10911944	DII	-0,03914	0,05015	-0,03539
rs12029788	DII	0,427	-0,1834	0,53419
rs4539083	DII	-0,27055	0,24979	-0,58606
rs12144629	DII	0,41014	-0,17364	-0,6142
rs10910079	DII	-0,59596	-0,20929	0,09884
rs713332	DII	-0,74582	-0,60228	-0,38571
rs7551490	DII	0,12685	-0,54474	-0,14735
rs16838636	DII	0,66714	0,35534	-0,00703
rs17403822	DII	-0,14605	-0,4286	0,20413
rs16858150	GLS	-0,02997	0,39953	0,10356
rs515279	GLS	-0,06199	-0,14684	0,00743
rs11588580	GLS	-0,15787	-0,26565	-0,7256
rs12143255	GLS	0,40156	0,21226	0,48673
rs549458	GLS	-0,55262	0,59403	-0,50197
rs4268311	GLS	0,05699	0,63281	-0,55058
rs16854997	GLS	-0,84944	-0,37796	0,03501
rs10799245	GLS	0,19611	-0,47426	0,15212
rs1608573	GLS	0,81279	-0,25583	-0,33379
rs12129643	GLS	-0,3282	0,29667	0,6483
rs3753758	HIP	0,85566	-0,36436	0,21175
rs1934120	HIP	0,03825	0,60336	0,00125
rs10911364	HIP	0,16013	0,23176	0,7186
rs2281173	HIP	1,1202	0,29145	-0,15326
rs11577132	HIP	-0,02099	-0,90937	0,0461
rs2245988	HIP	0,27608	-0,29682	-0,84937
rs11573997	HIP	-0,64117	-0,04222	-0,38703
rs1577592	HIP	-0,43267	0,00442	0,45967
rs7530685	HIP	0,02925	0,00764	-0,00849
rs2765033	HIP	-0,40274	-0,75116	0,51138

# Bibliografía

- [1] BALDING, David J.: A tutorial on statistical methods for population association studies. En: *Genetics* 21 (2006), Nr. 1, p. 195–203
- [2] BISHOP, Christopher M.: *Pattern Recognition and Machine Learning*. 1. Springer, 2006
- [3] CACÉRES, Abdiel: La métrica de Levenshtein. En: *Revista de Ciencias Básicas UJAT* 7 (2008), Nr. 2, p. 35–43
- [4] CHUANG, Li-Yeh ; HOU, Yu-Jen ; YANG, Cheng-Hong: A Novel Prediction Method for Tag SNP Selection using Genetic Algorithm based on KNN. En: *International Journal of Chemical and Biological Engineering* 3 (2010), Nr. 1, p. 12–17
- [5] DÍAZ, Luis.G: *Estadística Multivariada: Inferencia y Métodos*. 1. Universidad Nacional de Colombia : Facultad de Ciencias, 2002
- [6] GUERRERO, Flor ; RAMÍREZ, José: *El análisis de escalamiento multidimensional: Una alternativa y un complemento a otras técnicas multivariantes..* – Departamento de Economía y Empresa, Universidad Pablo de Olavide
- [7] GWAS: *Genome-Wide Association Study*, 2008
- [8] HALPERIN, Eran ; KIMME, Gad ; SHAMIR, Ron: Tag SNP selection in genotype data for maximizing SNP prediction accuracy. En: *Bioinformatics* 21 (2005), Nr. 1, p. 195–203
- [9] HAPMAP PROJECT: *International HapMap Project*, 2002
- [10] HASTIE, Trevor ; TIBSHIRANI, Robert ; FRIEDMAN, Jerome: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. Springer, 2009
- [11] HE, Jingwu ; ZHANG, Jun ; ALTUN, Gulsah ; ZELIKOVSKY, Alexander ; ZHANG, Yan-qing: Haplotype Tagging using Support Vector Machines. / Department of Computer Science, Georgia State University,. 2004. – Informe de Investigación
- [12] HE, Jinqwu ; ZELIKOVSKY, Alexander: Informative SNP Selection Methods Based on SNP Prediction. En: *IEEE* 22 (2006), Nr. 25, p. 1–8

- 
- [13] HE, Jinqwu ; ZELIKOVSKY, Alexander: MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. En: *Bioinformatics* 22 (2006), Nr. 20, p. 2558–2561
- [14] KRUSKAL, J.B: Nonmetric Multidimensional Scaling: A Numerical Method. En: *Psychometrika* 2 (1964), Nr. 2, p. 115–129
- [15] LEVENSHTAIN, V.I: *Binary codes capable of correcting deletions, insertions, and reversals*. 10. Soviet Union : Soviet Physics Doklady, 1966
- [16] LOPEZ, E ; HIDALGO, R: Escalamiento Multidimensional No Métrico. Un ejemplo con R empleando el algoritmo SMACOF. En: *Estudios sobre educación* 18 (2010), Nr. 18, p. 9–35
- [17] NOWOTNY, Petra ; KWON, Jennifer M. ; GOATE, Alison M.: SNP analysis to dissect human traits. En: *Current Opinion in Neurobiology* 11 (2010), Nr. 5, p. 637–641
- [18] R DEVELOPMENT CORE TEAM: *R: A Language and Environment for Statistical Computing*, 2009
- [19] SADAVA D, Heller C Hillis D Purves W.: *Life: Science of Biology*. 1. W. H. Freeman, 2006
- [20] SHEPARD, R.N: The analysis of proximities: multidimensional scaling with an unknown distance function. En: *Psychometrika* 27 (1962), Nr. 27, p. 125–140, 219–246
- [21] WANG, William Y. S. ; BARRATT, Bryan J. ; CLAYTON, David G. ; TODD, John A.: Genome-Wide Association Studies: Theoretical and Practical Concerns. En: *Nature Reviews* 6 (2005), Nr. 1, p. 109–118
- [22] WEEDON, Michael N. ; LANGO, Hana ; LINDGREN, Cecilia M. ; WALLACE, Chris: Genome-wide association analysis identifies 20 loci that influence adult height. En: *Nature Genetics* (2008), Nr. 40, p. 575–583