ANÁLISIS COMPARATIVO DE MULTIFRACTALIDAD EN GENOMAS HUMANOS COLOMBIANOS RELACIONADOS CON POBLACIONES HUMANAS

CHRISTIAN GUSTAVO ARIAS IRAGORRI

TESIS DOCTORAL

UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

SANTIAGO DE CALI
2019

ANÁLISIS COMPARATIVO DE MULTIFRACTALIDAD EN GENOMAS HUMANOS COLOMBIANOS RELACIONADOS CON POBLACIONES HUMANAS

CHRISTIAN GUSTAVO ARIAS IRAGORRI

TRABAJO DE GRADO PARA OPTAR EL TÍTULO: DOCTOR EN INGENIERÍA CON ÉNFASIS EN CIENCIAS DE LA COMPUTACIÓN

DIRECTOR: PHD. PEDRO ANTONIO MORENO TOVAR

UNIVERSIDAD DEL VALLE
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PLAN: 9702

SANTIAGO DE CALI

2019

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS, GRÁFICOS, ILUSTRACIONES Y FIGURAS

LISTA DE FIGURAS

LISTA DE TABLAS

Dedicado a:

Mi madre:

Por su amor y dedicación con sus hijos

Mi esposa:

Por su apoyo incondicional

A las 3M:

Por esta vida maravillosa

RESUMEN

El estudio de las secuencias genéticas y su relación con los diversos fenómenos biológicos presenta una gran relevancia ya que los resultados obtenidos pueden ayudar a la detección y prevención de enfermedades, mejoras en la productividad y resistencia de los cultivos, y un sinnúmero de aplicaciones. Debido a la complejidad y el volumen de datos disponibles además de la gran variedad de tipos de análisis posibles, especialmente en temas de alineamiento y búsqueda de patrones de secuencias, por lo cual, se hace necesario y urgente implementar nuevos abordajes que permitan desarrollar diagnósticos más precisos en los diversos campos y la búsqueda y comprensión de nuevos conocimientos. La mayoría de los algoritmos usados para estos estudios hacen comparaciones que buscan similitudes entre las cadenas que representan el ADN, usando algoritmos como la distancia Levenshtein y Jaro-Winkler, sin embargo, existe otro tipo abordaje donde las cadenas independientes dejan de ser el eje central de las comparaciones y lo que se busca en su lugar, son patrones estructurales, para lo cual se utilizan aproximaciones no lineales.

Similitudes lineales		**Similitudes estructurales**	
CCCTACC	CCCTACG	AACC…AACC…AACC	AACC..AACC
CCCTACC	CCTACG	AACCCAA	TTCCCTT
CCCTACC	CCCCTACC	ACCA	AACCCCAA

Debido a la complejidad y el volumen de datos disponibles además de la gran variedad de tipos de análisis posibles, se ha fomentado en la investigación el uso de diversos enfoques y técnicas, pasando por áreas tan diversas como técnicas de aprendizaje de máquinas, redes neuronales, modelos de Markov, algoritmos genéticos y fractales (por mencionar algunas), las cuales son potenciadas, gracias a los avances en los sistemas de cómputo, las bases de datos y los algoritmos. Por ejemplo, en la **genómica comparada** se usa una gran variedad de herramientas para comparar las secuencias de genomas completos de distintas especies, incluido el genoma humano.

El propósito de este trabajo en primer lugar, es realizar una comparación entre genes de diferentes organismos, utilizando enfoques no lineales, aplicando conceptos como la ley de Zipf y la dimensión fractal y, en segundo lugar, hacer

comparación de genomas completos de poblaciones humanas, incluida una muestra de la población colombiana, utilizando técnicas de análisis multifractal.

El contenido de esta Tesis Doctoral inicia con una introducción mencionando la complejidad y necesidad del análisis de secuencias, luego se hacen breves definiciones sobre la genética, así como de sistemas dinámicos, fractales y multifractales, posteriormente se exponen algunos de los estándares en la representación de datos y las bases de datos referencia. Luego se presentan los materiales y métodos utilizados en ambos experimentos para finalmente presentar los resultados y conclusiones de las dos investigaciones efectuadas.

El código fuente, el modelo de datos de la base de datos, las instrucciones para acceder a la base de datos y demás material suplementario, se encuentra disponible en el siguiente enlace: https://bit.ly/2N0btPH

Palabras clave:
Genoma, gen interrumpido, exón, intrón, fractalidad, dimensión fractal, conteo de cajas, representación del juego del caos, complejidad, multifractalidad, poblaciones, ancestría, orden filogenético, KEGG, enfermedades.

INTRODUCCIÓN

Las características biológicas de los seres vivos están determinadas por su ADN, el cual contiene los genes y está conformado por una cadena donde cada elemento es una base nitrogenada. El ADN sirve como un molde que después de pasar por múltiples procesos genera productos como las proteínas.

El estudio del genoma es de vital importancia ya que ayuda a entender la constitución y el desarrollo de los seres vivos, de esta forma se puede, entre otras cosas, mejorar la aplicación de los medicamentos, combatir las enfermedades de manera más eficiente y aumentar la productividad de los cultivos.

Actualmente, los costos de secuenciación han disminuido y los métodos utilizados han ido mejorando en velocidad y calidad, de tal forma que los datos disponibles han aumentado vertiginosamente, pasando de proyectos como la secuenciación del genoma humano[1] a proyectos como los 1000 genomas[2]. Esta situación ha generado la necesidad de investigar en mecanismos de análisis que apoyen a los investigadores en el proceso de caracterizar y procesar dicha información.

En los últimos años se ha evidenciado un gran interés por el estudio de los mecanismos genéticos y su relación con el desarrollo de los seres vivos y las enfermedades como el Cáncer[1][2][3]. En esa misma vía, se ha tratado de establecer relaciones entre genes y enfermedades utilizando diversas técnicas, entre las cuales se destacan los estudios de asociación poblacional o grupos humanos.

El análisis de dichas estructuras representa un reto debido al gran volumen de información que se debe procesar y a la complejidad de los datos y el proceso mismo. La utilización de herramientas computacionales es común y aunque hay sendos desarrollos en las áreas de alineamiento, no hay plataformas públicas que faciliten el análisis de las secuencias mediante métodos no lineales.

El Grupo de Bioinformática de la Universidad del Valle, entre otras temáticas ha abordado el análisis fractal y multifractal [4], [5] de diversos organismos, ahora se busca identificar relaciones entre las características no lineales y las características biológicas de diversas especies, en particular su predisposición a ciertas enfermedades como el cáncer en poblaciones humanas.

[1] Archivo de información del Proyecto del Genoma Humano: web.ornl.gov/sci/techresources/Human_Genome/home.shtml
[2] IGSR: El recurso internacional de muestra de genoma: www.1000genomes.org/

OBJETIVOS

OBJETIVO GENERAL

Desarrollar un análisis comparativo entre los genomas humanos colombianos y los demás genomas del proyecto “1000 genomas humanos” efectuando caracterizaciones multifractales y buscando su relación con fenómenos biológicos.

OBJETIVOS ESPECIFICOS

- Desarrollar un componente de software que permita la fácil transformación de una secuencia en su equivalente RJC.

- Desarrollar un componente de software que permita calcular el espectro multifractal de una secuencia a partir de parámetros previamente definidos.

- Diseñar una base de datos para el almacenamiento y clasificación de la información.

- Calcular el espectro multifractal para las secuencias pertenecientes al proyecto 1000 genomas humanos.

- Realizar un análisis exploratorio de los datos obtenidos y en particular la comparación de los genomas colombianos con la muestra representativa de los diversos grupos humanos incluidos en el estudio.

1 MARCO TEORICO

1.1 BIOINFORMATICA

La bioinformática es un campo interdisciplinario que involucra principalmente biología molecular, genética, ciencias de la computación, matemáticas, estadística y procesamiento masivo de datos donde problemas biológicos son tratados desde un punto de vista computacional. Las principales dificultades radican en realizar un correcto modelado de los procesos biológicos y posteriormente inferir conclusiones a partir de los resultados obtenidos[6].

Una solución de bioinformática usualmente incluye los siguientes pasos: Recolección de datos biológicos y la generación de sus estadísticas, la construcción de un modelo computacional, solución del modelo computacional y la prueba y evaluación del algoritmo computacional[6].

Los análisis bioinformáticos implican pasar los datos a través de una serie de transformaciones, llamadas una tubería o un flujo de trabajo (*pipeline*). Por lo general, estas transformaciones se realizan mediante software de línea de comandos ejecutable de terceros escrito para sistemas operativos compatibles con Unix. El avance en la precisión y velocidad y la disminución de los costos en actividades de secuenciación ha hecho que éste disponible un gran volumen de secuencias (y datos) que hace necesario contar con estos flujos de trabajo para su procesamiento eficiente[7].

1.2 COMPLEJIDAD Y SISTEMAS DINAMICOS NO LINEALES

Hay muchas definiciones de la complejidad, entre las cuales encontramos [8][9].La complejidad es una medida que se le puede aplicar a un sistema, el primer factor a tener en cuenta es la cantidad de elementos. Se puede decir que una ciudad es más compleja que un barrio ya tiene muchos más elementos interactuando en diversas escalas. El segundo factor es el grado de conectividad en el sistema, de hecho, cuando se tiene un bajo nivel de conectividad se puede describir el sistema enumerando las propiedades de los elementos individuales, pero cuando se incrementa la conectividad, también aumentan las relaciones que definen el sistema, por esto los sistemas complejos son típicamente modelados como redes que pueden capturar y cuantificar la información acerca de las relaciones entre sus

elementos. El tercer factor es la adaptación, lo que significa que los elementos se vuelven capaces de adaptar su comportamiento al tiempo que la incrementa la complejidad. Esta capacidad de adaptación también significa que los elementos pueden autoorganizarse, limitando la necesidad de un control centralizado, propiciando la aparición de organización en dichos elementos para interactuar y sincronizarse formando patrones. El cuarto factor es la diversidad, es decir que entre más grande sea la diversidad entre las partes, más complejo y abstracto debe ser un modelo para capturar las características comunes subyacentes.

En consecuencia, de lo anterior, se puede definir un sistema complejo como aquel que está compuesto de múltiples elementos los cuales están altamente interconectados y adaptados para lograr un propósito u ofrecer una funcionalidad. Los sistemas complejos pueden exhibir extraordinaria robustez, y en ocasiones extraordinaria fragilidad, donde algunos eventos a pequeña escala pueden tener un gran efecto sistémico, conocido popularmente como el efecto mariposa.

1.2.1 SISTEMAS DINÁMICOS

Un sistema dinámico es aquel que evoluciona en el tiempo, es decir, tiene reglas que hace que el estado presente sea determinado por estados anteriores. Se distingue entre *sistemas dinámicos lineales* y *sistemas dinámicos no lineales*. En los sistemas dinámicos lineales, existe una relación directa de causa efecto entre los elementos, por ejemplo, si X cantidad de gasolina me permite viajar una distancia Y, el doble de gasolina me permitirá viajar el doble de distancia. Desde el punto de vista matemático el segundo miembro de la ecuación es una expresión que depende en forma lineal de x[10], tal como: $X_{n+1} = 3X_n$. Si se conocen dos soluciones para un sistema lineal, la suma de ellas es también una solución; esto se conoce como *principio de superposición*. En general, las soluciones provenientes de un espacio vectorial permiten el uso del álgebra lineal y simplifican significativamente el análisis.

Los **sistemas no lineales** son mucho más difíciles de analizar ya que el principio de superposición no es aplicable a ellos, por ejemplo, si hay dos canciones que son del agrado de una persona, poner las dos al mismo tiempo generará interferencia entre ellas y como resultado no habrá más disfrute que escuchar cada una independientemente. La no linealidad emerge cuando existen algunas

relaciones entre los elementos del sistema que pueden ser sinérgicos, haciendo que la salida del sistema sea mayor a la suma de sus partes, o que, por el contrario, la interferencia haga que el resultado sea menor que la suma de los componentes individuales. La no linealidad también surge de ciclos de retroalimentación (*feedback loops*) donde el resultado de una iteración se utiliza como entrada para la iteración siguiente, por ejemplo, un sistema de interés compuesto, donde el capital del próximo periodo se calcula sobre el capital anterior más los intereses generados. Las funciones iterativas como la anterior son un concepto importante en la no linealidad y son un componente de la geometría fractal (sección 1.3.3).

Los sistemas dinámicos no lineales a menudo exhiben un fenómeno conocido como caos, con comportamientos totalmente impredecibles debido principalmente a la intrincada relación entre sus componentes y la gran sensibilidad al valor de sus variables de entrada.

1.2.2 TEORÍA DEL CAOS

Es la rama de las matemáticas, la física y otras ciencias (biología, meteorología, economía, entre otras) que trata ciertos tipos de sistemas complejos y sistemas dinámicos no lineales muy sensibles a las variaciones en las condiciones iniciales. Pequeñas variaciones en dichas condiciones iniciales pueden implicar grandes diferencias en el comportamiento futuro, imposibilitando la predicción a largo plazo. Estos sistemas son en rigor deterministas, es decir; su comportamiento puede ser completamente determinado conociendo sus condiciones iniciales[11][12].

1.2.3 ATRACTOR

Es uno de los conceptos fundamentales del Caos, que se utiliza para representar la evolución en un sistema dinámico. Este tipo de representación ya había sido usado por Henri Pointcaré[13]. Dentro de los atractores aparece un tipo denominado atractores extraños.

Un atractor se llama extraño si tiene una estructura fractal. Este suele ser el caso cuando las dinámicas son caóticas, pero también existen extraños atractores no caóticos. Si un atractor extraño es caótico y exhibe una dependencia sensible de las condiciones iniciales, cualquiera de los dos puntos iniciales alternativos cercanos al atractor, después de cualquiera de varios números de iteraciones, conducirá a puntos que están arbitrariamente separados (sujetos a los límites del atractor), y después de cualquiera de varios otros números de iteraciones conducirán a puntos que están arbitrariamente juntos. Por lo tanto, **un sistema dinámico con un atractor caótico es localmente inestable pero globalmente estable**: una vez que algunas secuencias han ingresado al atractor, los puntos cercanos divergen entre sí, pero nunca se apartan del atractor[14].

1.2.4 ESTRUCTURAS DISIPATIVAS

Constituyen la aparición de estructuras coherentes, autoorganizadas en sistemas alejados del equilibrio. Se trata de un concepto de Ilya Prigogine[15], que recibió el premio nobel de química «por una gran contribución a la acertada extensión de la teoría termodinámica a sistemas alejados del equilibrio, que sólo pueden existir en conjunción con su entorno».

El término *estructura disipativa* busca representar la asociación de las ideas de orden y disipación. **El nuevo hecho fundamental es que la disipación de energía y de materia, que suele asociarse a la noción de pérdida y evolución hacia el desorden, se convierte, lejos del equilibrio, en fuente de orden**.

1.2.5 CRITICALIDAD AUTOORGANIZADA

Es un término usado en física para describir (clases de) sistemas dinámicos que tienen puntos críticos como un atractor en su evolución temporal. El comportamiento macroscópico de los sistemas con criticalidad autoorganizada exhibe invariancias de escalas espaciales y temporales típicas de una transición de fase, por ejemplo ruido 1/f. Los sistemas que exhiben criticalidad autoorganizada espontáneamente se mantienen cerca del punto crítico, de ahí su nombre. El concepto fue concebido originalmente por Per Bak, Chao Tang y Kurt Wiesenfeld ("BTW") en [16] y es considerado uno de los mecanismos que generan la complejidad que nos rodea. El concepto ha sido aplicado en campos muy diversos incluyendo geofísica, cosmología, evolución y ecología, economía,

gravedad cuántica, sociología, física solar, física de plasma, neurobiología y otros campos.

Los fenómenos críticos autoorganizados pueden observarse en sistemas en desequilibrio con muchos grados de libertad extendidos y con algún grado de no linealidad. Muchos ejemplos validando este concepto han sido identificados desde el artículo original, pero aún no hay acuerdo sobre las condiciones necesarias y suficientes para que un sistema exhiba criticalidad autoorganizada.

1.3 GEOMETRIA Y ANÁLISIS FRACTAL

1.3.1 LEY DE POTENCIA

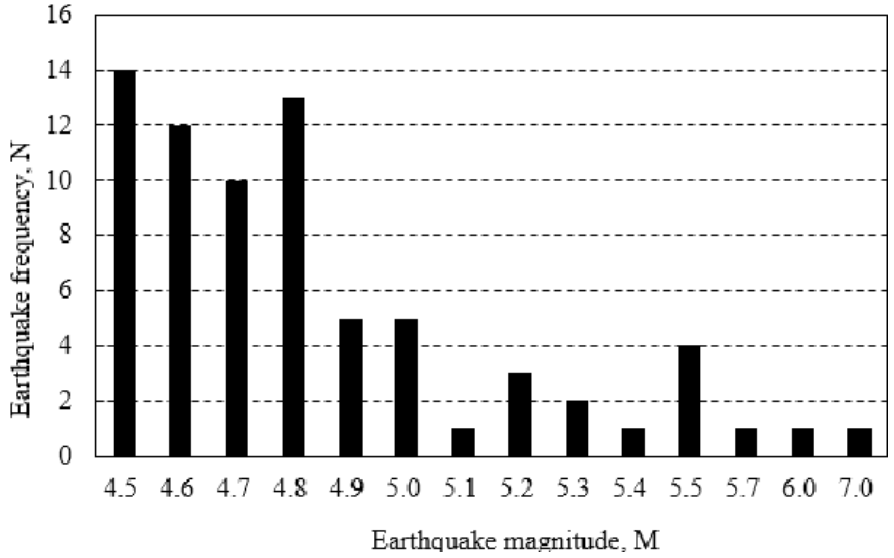
Es una relación matemática de dos magnitudes M y N descrita como:

$$M \sim C N^{p}$$

Donde C es un número real y P un exponente. Múltiples fenómenos siguen este tipo de relación donde las frecuencias son proporcionales al valor de las variables elevadas a un exponente constante.

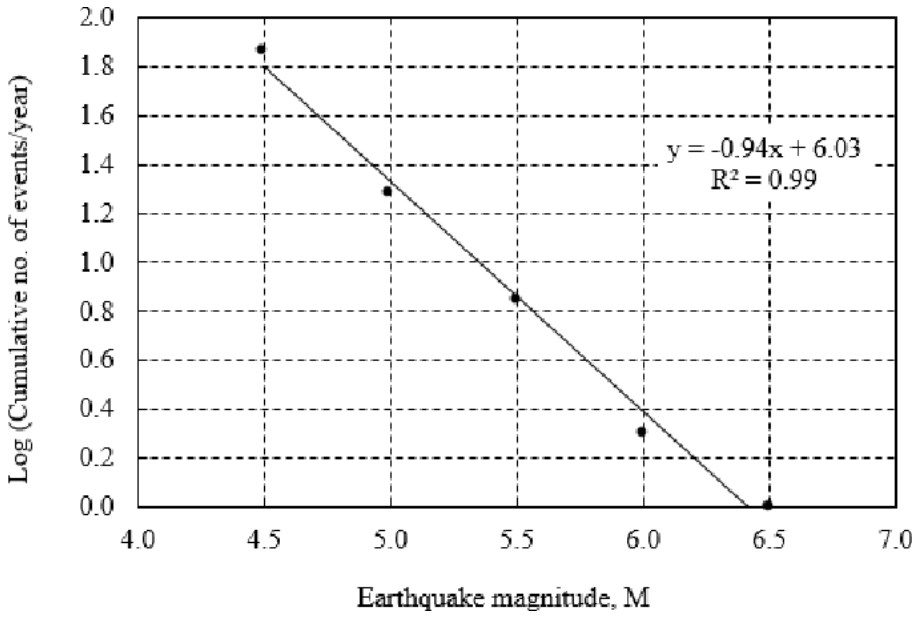
Muchos sistemas tienen propiedades que siguen la ley de potencia, un ejemplo particular de aplicación es una relación inversa, la cual tiene dimensión negativa, por ejemplo, la frecuencia de los terremotos es inversamente proporcional a su magnitud.

Figura 1: Histograma de frecuencia vs magnitud de terremotos producidos por las fallas de Khoy[17]



Una ley de potencia puede ser convertido en una relación lineal si las variables involucradas se grafican como logaritmos. Graficando las dos variables se puede determinar que dicha relación efectivamente se ajusta a la ley de potencias si el resultado tiende a ser una recta.

Figura 2: Log(Cantidad de eventos por año) vs Log(magnitud) de terremotos [17]



La importancia de la ley de potencia radica en que ayuda a evidenciar una regularidad subyacente en las propiedades de los sistemas estudiados. A menudo

sistemas altamente complejos tienen propiedades donde los cambios entre los fenómenos a diferentes escalas son independientes de la escala particular que se está observando. La imagen que tomamos a una escala es similar a la imagen que se observa en otra escala. Esta propiedad de autosimilitud es característica de relaciones de leyes de potencia. [18]

1.3.2 LEY DE ZIPF-MANDELBROT

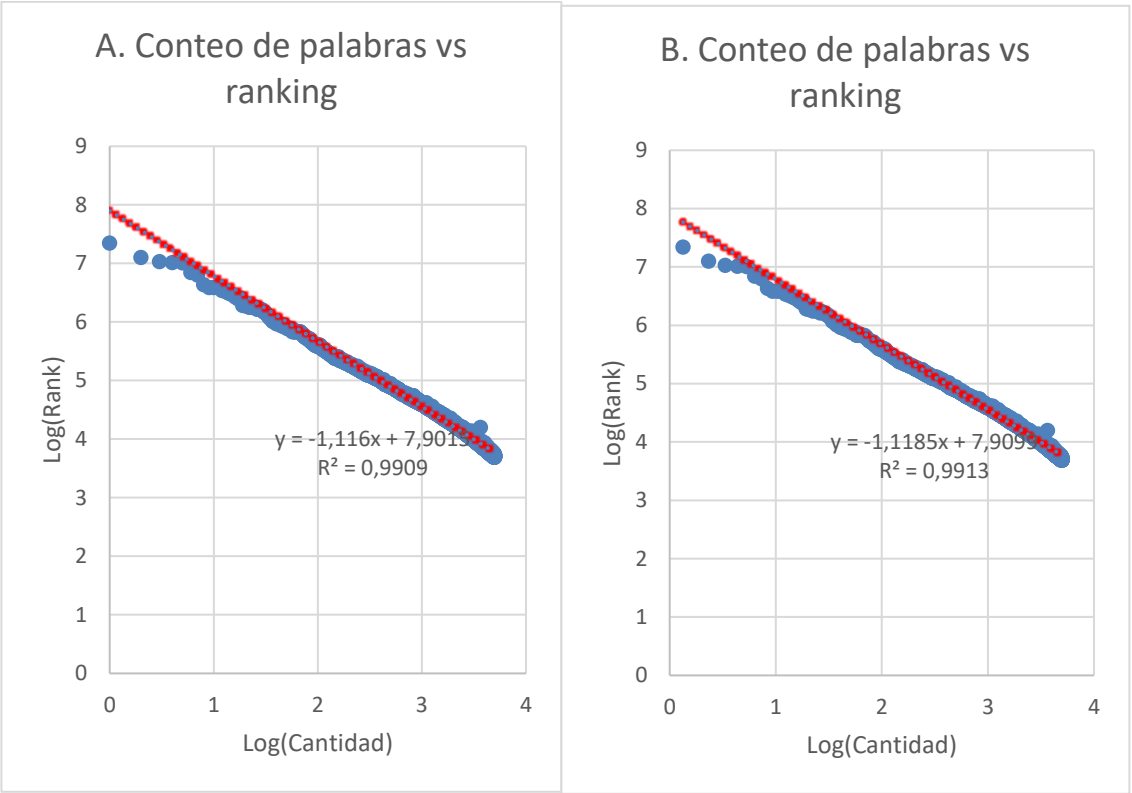
Inicialmente, la ley de Zipf fue formulada en la década de los 40s por George K. Zipfs, lingüista de la Universidad de Harvard, para el estudio de la frecuencia del uso de las palabras en un texto o discurso. Se cataloga como una ley de potencia definida por: $P_n \sim 1/n^a$, donde n representa la jerarquía del fenómeno observado en relación con otros fenómenos, por ejemplo: el más utilizado tendría un valor n=1, el segundo tendría un valor n=2 y así sucesivamente. α es un valor real positivo, generalmente mayor que uno; y P representa la frecuencia. En términos generales, esta ecuación indica que el segundo elemento ocurre aproximadamente a la mitad de la frecuencia del primer elemento y el n-ésimo elemento ocurre a una frecuencia cercana a 1/n [19].

Muchos sistemas siguen dicha ley, una de las áreas donde más se aplicado, es el estudio del lenguaje, según Mandelbrot, existen dos métodos para aplicar la ley de Zipf a un lenguaje natural: el primer método cuenta la frecuencia de las palabras utilizadas en un texto, mientras que el segundo método utiliza la longitud de las palabras utilizadas en dicho texto[20][21]. Ambos coeficientes de determinación pueden ser mejorados, al añadir factores de corrección cuando éstos son comparados con la ley de Zipf clásica [22] . Este enfoque es conocido como la ley de potencia de Zipf/Mandelbrot (en adelante Z/M) [19].

$$P_n \sim 1/n^{a+v}$$

En la Figura 3 se observa como la ley de Z/M tiene un mejor ajuste que la ley de Zipf original.

Figura 3: A. Aplicación de la ley de ZiPF al ranking y conteo de las 5000 palabras más frecuentes en el idioma ingles B. Aplicación de la ley de Z/M. A pesar de que en esta correlación el valor es alto, se logra apreciar el mejor valor de R^2 en el lado 3B.



En el primer método, la frecuencia de las palabras es estimada y ordenada por jerarquías, de la más frecuente a la menos frecuente. Posteriormente se calcula una ley de potencia entre la jerarquía y la frecuencia de uso de las palabras, donde el exponente escalar determina la complejidad o riqueza del vocabulario presente en el texto. El segundo enfoque, utiliza la longitud de las palabras, en este contexto, cuando la frecuencia de longitud de las palabras es ajustada a una distribución de Zipf, el exponente de potencia puede variar significativamente de 1, con una desviación dependiente de la riqueza del vocabulario presente en el texto [19].

1.3.3 FRACTALES

La geometría tradicional (euclídea) se encarga de las propiedades y de las mediciones de objetos tales como puntos, líneas, planos y volúmenes. A diferencia de estos objetos geométricos clásicos, que poseen propiedades de continuidad y diferenciabilidad, existen otros objetos geométricos irregulares que presentan estructura a cualquier escala y tienen un número infinito de singularidades (puntos

no diferenciables). Ejemplos de estos son las formas encontradas en la naturaleza, como montañas, franjas costeras, sistemas hidrográficos, nubes, hojas, árboles, vegetales, copos de nieve, y un sinnúmero de otros objetos que no son fácilmente descritos por la geometría tradicional. La geometría fractal provee una descripción matemática de estas formas irregulares que se denominan fractales. El termino fractal (del latín fractus – fracturado-) fue propuesto por Benoit Mandelbrot 1975 para referirse a objetos muy irregulares que no se podían describir en términos geométricos tradicionales y que además presentaba características de autosilimilitud, es decir que presentaba similitud de forma en diferentes escalas[22], [23].

Existen básicamente tres tipos de estructuras fractales en la naturaleza. Los fractales naturales como la forma de las costas de los continentes, las montañas, los cráteres de la luna, entre otras. Los fractales artificiales, como el set de Mandelbrot o el set de Julia, entre otros, los cuales son generados por Sistemas de Funciones Iteradas en el computador y 3) los fractales estadísticos, que están en los datos y sólo se aprecian cuando éstos son analizados por leyes de invarianza de escala través de la dimensión fractal (D) (

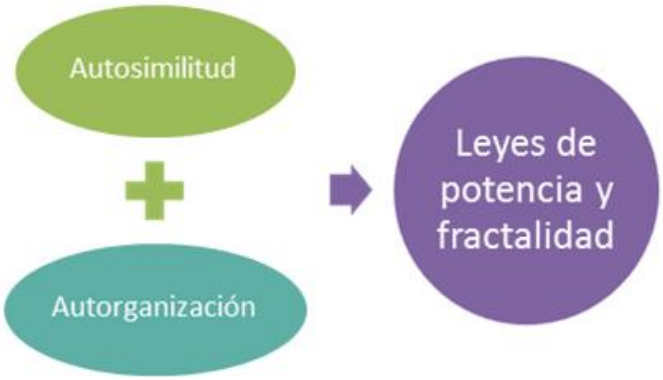
).

Figura 4: Fractales natural, artificial y estadístico.

Los objetos fractales tienen propiedades muy particulares, como la auto semejanza y la apariencia irregular, que permiten caracterizarlos en base a medidas cuantitativas relativas a su grado de irregularidad. Sean naturales o geométricos, todos los fractales poseen dimensiones fraccionarias. No coinciden, en general, con las dimensiones euclidianas que se miden con números enteros (0

es la dimensión de un punto, 1 es la dimensión de una línea, 2 es la dimensión de una figura plana, 3 es la dimensión de un cuerpo espacial). Esta dimensión fractal indica el grado según el cual el objeto fractal llena la dimensión euclidiana en la que está inmerso, esto es, la dimensión fractal es una medida de la irregularidad. Un fractal natural de dimensión 2.8, por ejemplo, tendría una forma de esponja aproximadamente tridimensional en apariencia. En cambio, un fractal natural de dimensión 2.2, sería un objeto mucho más suave que estaría muy cerca de parecer plano.[24]. La dimensión fractal da cuenta de cuán completamente parece llenar un fractal el espacio conforme se amplía el primero hacia escalas más y más finas y su cálculo se utiliza para conocer el grado de “fracturación” del sistema cuyas implicaciones dependen de la naturaleza de este.

Figura 5: Autosimilitud, autoorganización y leyes de potencia y fractalidad.



La dimensión fractal puede ser calculada mediante diversas técnicas dentro de las que se destaca el llamado “conteo de cajas”, el cual consiste en dividir una imagen aplicando una cuadricula con cajas de tamaño T, posteriormente se cuenta la cantidad de cajas N que se requieren para cumplir completamente la imagen. Luego se varía el tamaño de las cajas y se repite el proceso. La dimensión fractal será entonces la pendiente de la línea resultante del log(N) contra el log(T):

$$D = \frac{\log(N)}{\log(1/T)}$$

Así como se habla de “objetos fractales” se puede hablar de la “fractalidad” de un conjunto de datos, donde se analiza su periodicidad y autosemejanza a diferentes escalas birndando un entendimiento “estructural” del fenómeno analizado.

Una secuencia de ADN representa una serie de datos caracterizables por su dimensión fractal, dicha caracterización ha sido utilizada en estudios como la

diferenciación de secuencia codificantes y no codificantes[25], el análisis de propiedades autosimilares en la red neural del *Caenorhabditis elegans*[26], propiedades fractales del genoma humano[27] y otros.

1.3.4 REPRESENTACION DEL JUEGO DEL CAOS (RJC)

La representación del juego del caos consiste en una representación gráfica de una secuencia de valores. Brinda la posibilidad de transformar una secuencia de valores discretos en valores numéricos, los cuales son utilizados posteriormente para realizar diversos tipos de análisis. La idea general consiste en dibujar en un cuadrado una unidad donde cada arista representa un posible valor de la secuencia (suponiendo que hay 4 posibles valores). El procedimiento procesará cada uno de los valores de la secuencia. Inicialmente se ubica un punto en la mitad entre el centro del cuadrado y la arista que representa el primer valor de la secuencia, para los demás valores se ubican puntos en la mitad entre el punto anterior y la arista que representa el valor de la secuencia. Las coordenadas de dichos puntos son la representación numérica de la secuencia original. Existen generalizaciones para procesar secuencias con más de cuatro valores posibles como el Universal Sequence Map (USM)[28]. Las representaciones numéricas de las secuencias sirven como insumo para realizar caracterizaciones de secuencias que servirán posteriormente para efectuar procedimientos como el cálculo de su dimensión fractal. Este enfoque ha sido utilizado en diversos estudios del ADN como en [29]–[32].

1.3.5 ANÁLISIS MULTIFRACTAL

Su propósito es caracterizar sistemas dinámicos por medio de una función denominada espectro multifractal. Es una generalización del análisis fractal donde una sola dimensión fractal no es suficiente para describir la dinámica del sistema, y en su lugar se utiliza un espectro de valores[33]. Para ello se usan múltiples reglas es escalamiento hasta obtener un espectro de dimensiones D_q, para todos los q (o momentos) que son evaluados:

$$D_q(\varepsilon) = \frac{log\left(\sum_i \left(\frac{M_i}{M_0}\right)^q\right)}{\log(\varepsilon)\,(q-1)}$$

Donde M_i es la cantidad de puntos que caen en la i^{esima} caja y M_0 es la cantidad total de puntos (M_i/M_0 sería el porcentaje de puntos que caen la caja), ε es el tamaño de la caja dado por 1/E donde E es la cantidad cajas (también llamado factor de escalamiento). Se debe tener en cuenta una consideración especial para el caso en q = 1 donde el denominador debe ser reemplazado por $log(1/\varepsilon)$.

El análisis multifractal ha sido utilizado con éxito para el análisis de secuencias como en [4], [5], entre otros trabajos.

1.3.6 APLICACIÓN DE ANÁLISIS NO LINEALES

Se han intentado muchas y diversas aproximaciones para el estudio de los genomas, en particular se han recurrido a métodos matemáticos aplicados a otros campos como la física como se ilustra en[34]–[36]. Ya que las secuencias genómicas pueden tratarse como ondas o señales, diferentes técnicas y aproximaciones no lineales han sido utilizadas para su análisis. Se ha utilizado la teoría del juego del caos principalmente para hacer comparaciones de patrones, análisis de frecuencias búsquedas de motivos y representaciones graficas de ADN como se refleja en [28]–[30], [32], [37]–[41], los fractales se han utilizado para analizar la riqueza de información de diversas zonas y el análisis de estructuras y patrones mediante el cálculo de la dimensión fractal de diversos elementos como en [26], [40], [42]–[44], los multifractales se han utilizado para caracterizar secuencias mediante el cálculo del espectro multifractal como en estudio de la multifractalidad del genoma humano [5] y el *Caenorhabditis elegans* [4].

1.4 GENÉTICA

La genética (del griego genetikos) es el estudio de la herencia biológica que se transmite de generación en generación por medio del genoma. El genoma contiene la información biológica necesaria para construir y mantener un ejemplo viviente del organismo[45].

La definición de gen más actual ha logrado gran consenso alrededor de: “una región localizable de una secuencia genómica, correspondiente a una unidad de herencia, la cual es asociada a regiones regulatorias, regiones transcritas y otras regiones funcionales”[46], [47].

Al conjunto de genes y regiones intergénicas contenidos en un cromosoma se le denomina genoma, el cual está compuesto por Acido Desoxirribonucleico (ADN), el cual es un polímero, cuyas subunidadades monoméricas son cuatro nucleótidos (adenina, guanina, citosina y timina) que se pueden unir en cualquier orden formando cadenas desde cientos hasta millones de unidades de longitud[45]. Estas cadenas sufren una serie de procesos a lo largo de la ecuación del flujo de la información genética tras el cual se producen proteínas y otras substancias. Dicho proceso se conoce como el dogma central de la biología molecular y fue propuesto inicialmente por Crick en [48] y paulatinamente ha sido complementada. Uno de estos procesos, la **replicación del ADN** es el responsable de heredar la información genética a las células hijas o a la descendencia, a través de las divisiones celulares (fisión binaria (procariotes), mitosis y meiosis (eucariotas)).

Por otra parte, los mapas genéticos (mapas de genomas y cromosomas) son representaciones de las secuencias genómicas que se pretenden estudiar. En su forma más básica es simplemente una secuencia de caracteres que representan los diferentes nucleótidos y sus posiciones y en variantes más complejas se puede llegar a asociar información de los genes, sus posiciones, variantes, y muchas características adicionales. Dependiendo del tipo de estudio que se quiera realizar se escoge la representación más adecuada.

1.4.1 GENOMA HUMANO REFERENCIA

El genoma referencia es una secuencia de ADN que hace las veces de “muestra representativa” de una especie, de tal forma que, al codificar un organismo de la misma especie, se puede almacenar solo los segmentos diferentes al genoma referencia[49]. El tener un genoma referencia permite establecer una base contra la cual se pueden comparar diversos organismos. Los genomas referencias de las especies secuenciadas hasta ahora se encuentran depositados en los ficheros del NCBI (ftp://ftp.ncbi.nlm.nig.gov/) y replicados en otras bases de datos.

El genoma humano (Homo sapiens) contienen cerca 6.400 millones de pares de bases distribuidos en 23 pares de cromosomas (set diploide) que codifican cerca de 22.000 genes para proteínas con sus diferentes variantes alélicas. Periódicamente se publican nuevas versiones del genoma referencia acorde con la mejora de los métodos de secuenciación.

1.4.2 REPRESENTACION DE LA INFORMACIÓN GENETICA

Existe una gran variedad de formas para visualizar y representar la información genética. Su complejidad y contenido depende del propósito que se tenga para dicha información. Algunos simplemente representan la secuencia de una forma plana simplemente como una cadena de letras que representa las diversas bases (fasta) y otros incluyen información sobre el estudio que genero los datos, los autores, el inicio y fin de cada región y un sin número de etiquetas adicionales (GBK). A continuación, se presentan las descripciones del tipo de archivos más comunes y potencialmente útiles para este estudio. En la

se lista las referencias de otros formatos.

1.4.2.1 Archivos fasta

Son archivos de texto plano que por medio de una letra representan un nucleótido, también se puede usar para representar proteínas donde cada letra corresponde a un aminoácido. Permite incluir comentarios sobre la secuencia que se encuentra al interior del archivo. Por su simplicidad es ampliamente utilizado y se trabaja fácilmente con herramientas procesadoras de texto o con lenguajes de programación orientados a cadenas. La primera línea corresponde a la descripción de la secuencia e inicia con el símbolo '>'. Ejemplo de un archivo fasta:

>gi|3790755|gb|AF099922.1| Caenorhabditis elegans cosmid F56F11AGAAAATCAAGATGTTTCCAAAAATGACATCATGGTCGCTCAATCCAATAAGTATTAAGCCGCA
GCCACTGGTAGTCATGGCAAAAAGGAAACTCTTGTTGCGAATAGCATTTTTTCTATGAATCACCACGCAA
AAGATGATAACTGGCAAACCAATAATGCACATACAACAACCAGTAATGACCATAGCTTCAAAGATCCTGC

GATTGACAAATGTGTGAGAGCTGGACACTATCAGGGTGAGAGAGCAGCTTGT
ACCAGA

1.4.2.2 Archivos .GBK

El formato GenBank permite etiquetar diversos elementos de la secuencia. La cabecera contiene información que describe la secuencia como la fuente, la referencia y descripción. Posteriormente se encuentran las características del genoma incluyendo las proteínas que codifica y finalmente se encuentra la secuencia de ADN. Finaliza con los caracteres '//'. De él se pueden extraer diversa información como la posición, nombre y hebra en la que se encuentra un gen y otro tipo de características. Ejemplo del contenido de un archivo GBK:

```
LOCUS       NC_000908            580076 bp    DNA     circular BCT 12-NOV-2008
DEFINITION  Mycoplasma genitalium G37, complete genome.
ACCESSION   NC_000908
VERSION     NC_000908.2  GI:108885074
KEYWORDS    .
SOURCE      Mycoplasma genitalium G37
  ORGANISM  Mycoplasma genitalium G37
            Bacteria; Tenericutes; Mollicutes; Mycoplasmataceae; Mycoplasma.
REFERENCE   1  (bases 1 to 580076)
  AUTHORS   Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R.,
            Maruf,M., Hutchison,C.A., Smith,H.O. and Venter,J.C.
  TITLE     Essential genes of a minimal bacterium
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 103 (2), 425-430 (2006)
   PUBMED   16407165
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to
final
            NCBI review. The reference sequence was derived from L43967.
            On Jun 13, 2006 this sequence version replaced gi:12044850.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..580076
                     /organism="Mycoplasma genitalium G37"
                     /mol_type="genomic DNA"
                     /strain="G37"
                     /db_xref="taxon:243273"
```

```
     gene            686..1828
                     /gene="dnaN"
                     /locus_tag="MG_001"
…
ORIGIN
        1 taagttatta tttagttaat acttttaaca atattattaa ggtatttaaa aaatactatt
       61 atagtattta acatagttaa ataccttcct taatactgtt aaattatatt caatcaatac
…
   579961 atgatcctgc aacattagtt gccattgtag tttttaatac gccgccttta ttatttacaa
   580021 aagaaatgat catatattta aatgattata atatttcttt aatactaaaa aaatac
//
```

1.4.2.3 Archivos .VCF (Variant Call Format)

Estos archivos son utilizados principalmente para almacenar polimorfismos de nucleótidos simples (SNPs). Son archivos de texto plano que contiene líneas de metainformación, una línea de encabezado y finalmente los datos incluyendo su posición en el genoma. Este formato es importante porque sus datos pueden usarse para reemplazar los valores correspondientes en el genoma de referencia y de esta forma se evita la necesidad de transmitir el genoma completo de diversos individuos de la misma especie y en su lugar solo se transmitiría el genoma de referencia el VCF de cada uno. Este tipo de archivo es utilizado por el proyecto 1000 genomas humanos. Ejemplo de un archivo VCF:

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT QUAL FILTER INFO                              FORMAT
20     14370   rs6054257 G   A   29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ
```

Tabla 1: Referencias de otros tipos de archivos de secuencias

fastaq	http://maq.sourceforge.net/fastq.shtml
fastaqc	https://biof-edu.colorado.edu/videos/dowell-short-read-class/day-4/fastqc-manual

sam y bam	https://samtools.github.io/hts-specs/SAMv1.pdf
gff y gtf	http://useast.ensembl.org/info/website/upload/gff.html

1.4.3 GENES INTERRUMPIDOS (GI)

Los genes y genomas contienen la información de la vida orgánica. Están organizados en un mosaico de secuencias codificadas y no codificadas. En organismos eucariotas, los genes son llamados genes interrumpidos (GIs), los cuales son estructuras de series alternadas de intrones y exones a partir de las cuales se forma una transcripción primaria. Las secuencias derivadas de los intrones deben ser removidas para formar el ARNm maduro, el cual luego es transformado en proteína. Esta organización alternada sugiere una combinación aleatoria de intrones y exones y una organización modular de los productos proteicos. Actualmente, se ha secuenciados 500 genomas eucariotas y la longitud de los intrones y exones está siendo estudiada utilizando diferentes enfoques estadísticos, estructurales, funcionales y evolutivos. En muchos de ellos, las señales consenso, el contenido G+C y las distribuciones de longitud han sido caracterizadas para identificar algún grado de regularidad estructural. Estas características son de vital importancia para entender la arquitectura de los genes eucariotas, así como para desarrollar algoritmos de predicción genética.

De igual manera, los patrones comunes en la longitud de intrones y exones presentes en genes no relacionados sugieren su participación en el ensamblaje de genes y en el diseño y arquitectura del genoma humano [50]. Estas investigaciones revelan que la organización del gen interrumpido es una característica universal del gen eucariota que facilita la aparición de patrones alternados de splicing y patrones compartidos a lo largo de la escala filogenética.

1.4.4 ANCESTRÍA Y POBLACIONES HUMANAS

La ancestría es el estudio de los antepasados de un organismo. La información obtenida sirve para múltiples propósitos, en especial en estudios de ADN debido a la transferencia genética que se da de padres a hijos. Es así como se ha

identificado un gran número de enfermedades son hereditarias y que posibilidades tienen de manifestarse.

En un escenario más amplio, se ha tratado de establecer poblaciones humanas a partir de su ubicación geográfica y su ADN común. En el pasado, estas poblaciones presentaban menos mezclas con otros debido principalmente a las distancias existentes entre ellos (de todo tipo, no solo geográficas) y los incipientes medios de transporte, sin embargo, en la actualidad, debido a la globalización y mejora considerable en la posibilidad de desplazarse, estas mezclas son cada vez más comunes.

Se han establecido poblaciones humanas tipo a partir de las cuales se han basado numerosas investigaciones. Estas poblaciones están asociadas principalmente a su sitio de origen, es así como encontramos las denominaciones de africanos, amerindios, europeos, etc. Otros proyectos se han enfocado en caracterización de poblaciones como [51]–[53][54].

Hay sitios especializados para tratar de identificar la ancestría de las personas a través de un análisis genético como: 23andMe[3].

2 DIFERENCIACIÓN DE LA COMPLEJIDAD FRACTAL UTILIZANDO LA LEY DE ZIPF-MANDELBROT EN GENES INTERRUMPIDOS (GI)

2.1 INTRODUCCIÓN

El estudio de la longitud de los intrones y exones se ha hecho en el pasado por múltiples autores y en diversos organismos [50], [55]–[58] , generalmente desde el punto de vista estadístico, evolutivo, y relacionado con el proceso de splicing [59] o con la estructura de las proteínas[60]. Por ejemplo, Wang y Stein en [61] asumen que la evolución de la estructura exón/intrón es un proceso estocástico y que las características de este proceso pueden ser entendidas examinando su registro histórico, y la distribución de los tamaños actuales de los exones

[3] 23andMe: www.23andme.com/ancestry/

traducidos internamente. Estos autores proponen un proceso general de fragmentación aleatoria para caracterizar la dinámica evolutiva de la estructura exón/intrón.

A pesar de estas contribuciones, a la fecha no existe trabajo alguno de la longitud de los genes, exones e intrones abordado desde una perspectiva no lineal. A fin de tomar ventajas de las propiedades e implicaciones de la teoría de los sistemas dinámicos no lineales, en el presente trabajo se postula la hipótesis de que el tamaño de los genes o el proceso de fragmentación del GI (en exones e intrones) se puede examinar a través de una analogía lingüística utilizando la ley de potencia Z/M.

2.2 MATERIALES Y MÉTODOS

Ley de Zipf-Mandelbrot (Z/M)

La ley de Z/M: $P(S + V) = (S + V)^{-1/D}$ fue programada, donde P(S + V) es la frecuencia de las longitudes de todos los genes (o CDSs) de un organismo (enfoque A) o de los exones e intrones de un GI (enfoque B) y (S + V) es el rango más el parámetro V, siendo $V = 1/N-1$, y $N = 4$, ya que el alfabeto es dado por A, G, T y C. Posteriormente, la dimensión fractal (D) o valor de complejidad y el coeficiente de determinación, R^2 fueron calculados para cada grupo de CDSs/organismo o exones e intrones/GI, usando una distribución bilogarítmica.

Analogía lingüística Z/M

En la analogía lingüística Z/M se deben definir tres términos: el discurso, la palabra y el alfabeto. **1) el discurso o texto** sobre el cual se va a calcular la dimensión fractal, la cual nos da una medida del grado de riqueza o complejidad del texto. **2) la palabra,** es una unidad de información con definición y sentido lingüístico que existe en un idioma en particular, **3) el alfabeto**, es el conjunto de letras diferentes con las que se forman las palabras de un idioma en particular[20], [21]. Estos términos son aplicados a los textos lingüísticos cuando se desea estudiar la frecuencia del uso de las palabras y la riqueza del vocabulario empleado.

Para el presente trabajo, los términos de la analogía lingüística fueron cambiados y redefinidos a fin de ser aplicados al estudio de las longitudes de los genes (CDS), exones e intrones, (

).

Tabla 2: Términos de la analogía lingüística aplicados.

TÉRMINOS DE LA ANALOGIA LINGÜÍSTICA	**ENFOQUES**	
	A	**B**
Discurso	Set de genes por organismo	Gen interrumpido
Palabra	Longitud de cada gen/ organismo en pb	Longitud de cada intrón o exón/gen en pb
Alfabeto	A, G, T, C	A, G, T, C

En el **enfoque A**, la longitud de los GIs son las palabras, de modo que se obtiene un único componente escalar por organismo, lo cual nos daría una idea general del ajuste de la ley Z/M de la longitud de los genes. Esto significa que si tengo 228 organismos se van a obtener 228 dimensiones fractales.

En el **enfoque B**, se podría obtener un componente escalar por GI, a través de las longitudes de sus unidades de información (intrones y exones, en adelante UI). En el presente trabajo se analizaron 4’232.493 de genes eucariotes pertenecientes 219 organismos, lo que significa que se obtuvieron 4’232.493 de D. En ambos enfoques se utilizó el valor constante V = 1/3 con el que se obtiene el mejor ajuste posible.

¿Cómo interpretar el valor D de la analogía Z/M?

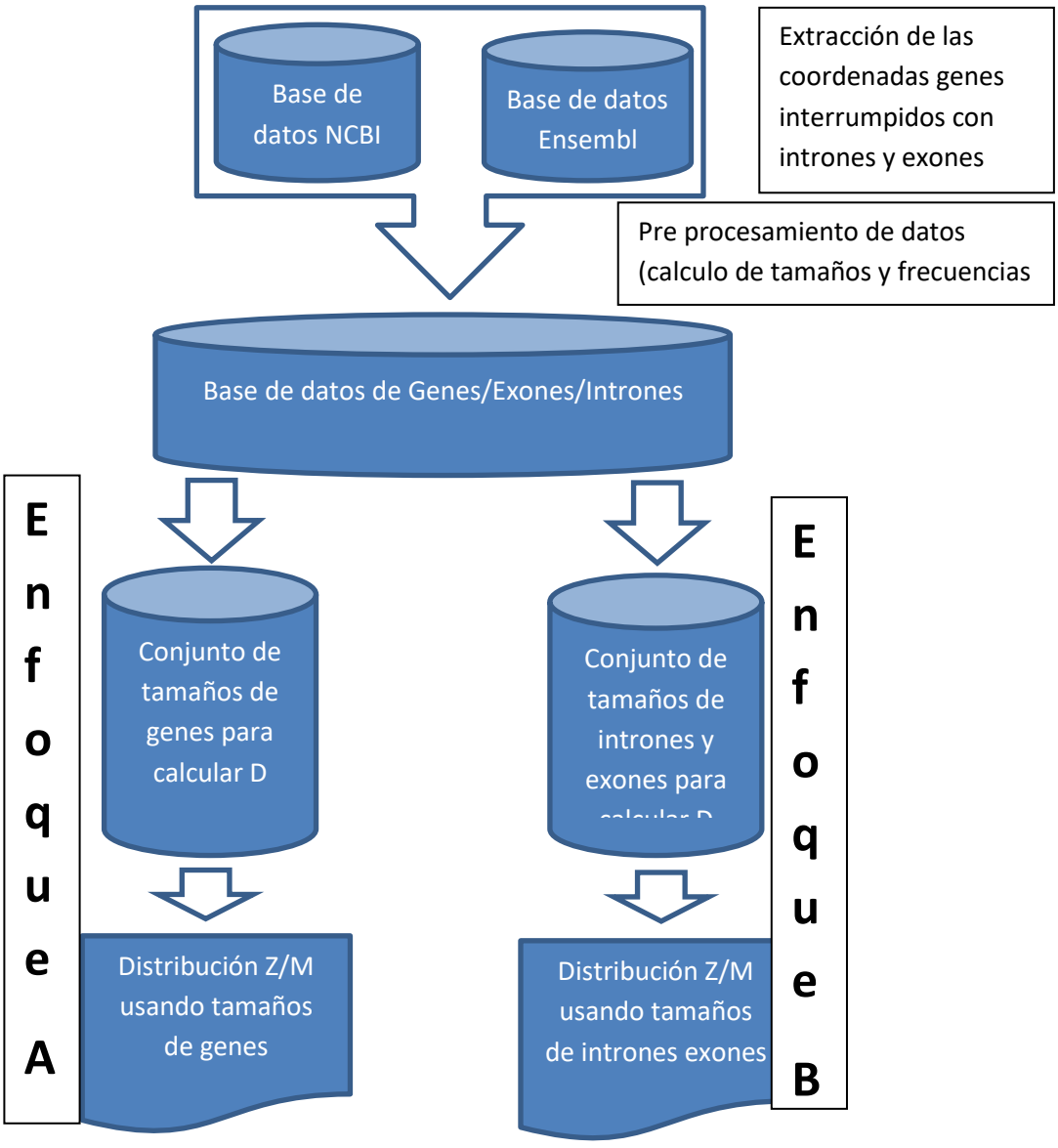
Según Mandelbrot, cuando un discurso tiene una dimensión fractal D < 1.0 se considera que el texto tiene una estructura fractal, mientras que aquellos textos con D > 1 se consideran texto contaminados o mutados, es decir, alterados con frecuencias de palabras en otro idioma. Igualmente, Mandelbrot considera que los textos tienen altas frecuencias de las palabras como los artículos y las preposiciones, las cuales, son conectores de las palabras que tienen bajas frecuencias y definen la gramaticalidad del texto y que dan sentido al discurso. Es en estos términos principalmente en que se interpretan los resultados de la

presente investigación, apoyados por evidencias de otros tipos de estudios y evidencias.

Bases de datos y herramientas

Las longitudes de los genes, exones e intrones se obtuvieron a partir de las bases de datos de 228 organismos depositadas del NCBI y Ensembl. Posteriormente se anotaron y tabularon las coordenadas de los genes codificantes para proteínas (enfoque A) y de los exones e intrones de los genes interrumpidos (enfoque B). Luego se calcularon los tamaños de los genes y de las UIs de cada gen interrumpido. Para ambos enfoques se utilizó la ley de Z/M a fin de obtener la dimensión fractal para cada organismo (enfoque A) y para cada gen interrumpido (enfoque B) (Figura 6).

Figura 6: Flujograma del análisis fractal de genes interrumpidos.



Para realizar el procesamiento de los datos se definió como mecanismo de persistencia una base de datos relacional, en la cual se almacenaron los genes y UIs con sus identificaciones, coordenadas, organismo, taxonomía y demás características de interés. También se almacenaron las frecuencias de los tamaños de dichas estructuras y los cálculos posteriores de las dimensiones fractales teniendo en cuenta el tamaño de las estructuras y su ranking.

Análisis de genes por agrupamiento jerárquico de exones e intrones por dimensión fractal y rutas metabólicas

Para el análisis de las D se utilizaron rangos de D entre 0 y 2. Para el análisis estructural y funcional de los genes y las rutas metabólicas relacionadas a los rangos fractales (D < 1.0) y no fractales (D > 1.0) se descargaron las rutas desde la base de datos de KEGG para crear un sistema de equivalencias de identificación de genes obtenidos previamente y las anotaciones de KEGG. Para visualizar los resultados se utilizaron **mapas de calor** a fin de determinar la relación entre las dimensiones fractales de los genes y las categorías y subcategorías de las rutas metabólicas obtenidas de KEGG.

En la siguiente tabla se encuentran las herramientas utilizadas para efectuar todos estos análisis.

Tabla 3: Herramientas informáticas utilizadas para la aplicación de la ley de Z/M.

Herramienta	Utilidad	
wget[4]	Herramienta de uso libre para descargar archivos mediante protocolos HTTP, HTTPS, FTP, FTPS	Se utilizó para realizar la descarga de los archivos de secuencias y archivos de bases de datos mysql
Mysql[5]	Software de base de	Se utilizó para restaurar las copias de las

[4] Sistema Operativo GNU: www.gnu.org/software/wget/
[5] Sistema de Gestión de Bases de datos: www.mysql.com/

	datos	bases de datos descargadas desde Ensembl que no estaban disponibles online.
Oracle[6]	Software de base de datos de la compañía Oracle	Se utilizó como repositorio de la base de datos donde se almacenaron las referencias a los genes, intrones, exones y rutas metabólicas de los diversos organismos
Pentaho Data Integration CE[7]	Software para realizar procesos de ETL (Extract, transform and load). Se utilizó la versión comunitaria que es de uso gratuito.	Se utilizó para descargar y transformar información desde las bases de datos de Ensembl que estaban en mysql y almacenarlas en la base de datos Oracle, también se manejó para automatizar algunos de los pipelines al integrar diversas fuentes de datos, adicionalmente se utilizó para establecer la asociación entre los genes y rutas metabólicas de KEGG
Eclipse[8]	Entorno integrado de desarrollo multilenguaje	En dicha herramienta se realizó la programación de las rutinas utilizadas para aplicar la ley de Z/M y así encontrar la fractalidad de los organismos y los genes interrumpidos.
Java[9]	Lenguaje de programación multipropósito orientada a objetos	Sobre este lenguaje se implementaron las rutinas de cálculo de la ley de Z/M.
PowerBI[10]	Herramienta de Microsoft para la generación de reportes	Con esta herramienta se generaron algunas de las figuras utilizadas en el presente trabajo.
Origenlab[11]	Herramienta para elaborar mapas de calor	Permite identificar cuáles son los puntos que centran la atención del usuario, a través de interacciones y pautas de

[6] Oráculo Aplicaciones Integradas en la nube y servicios de plataforma: www.oracle.com/index.html
[7] HITACHI: https://community.hitachivantara.com/s/pentaho
[8] Innovación y Colaboración La Fundación Eclipse: www.eclipse.org/
[9] JAVA: www.java.com/es/
[10] Power *Bi*: https://powerbi.microsoft.com/es-es/desktop/
[11] Origin Lab: www.originlab.com/

		comportamiento preestablecidas (entre atributos y categorías).

2.3 RESULTADOS

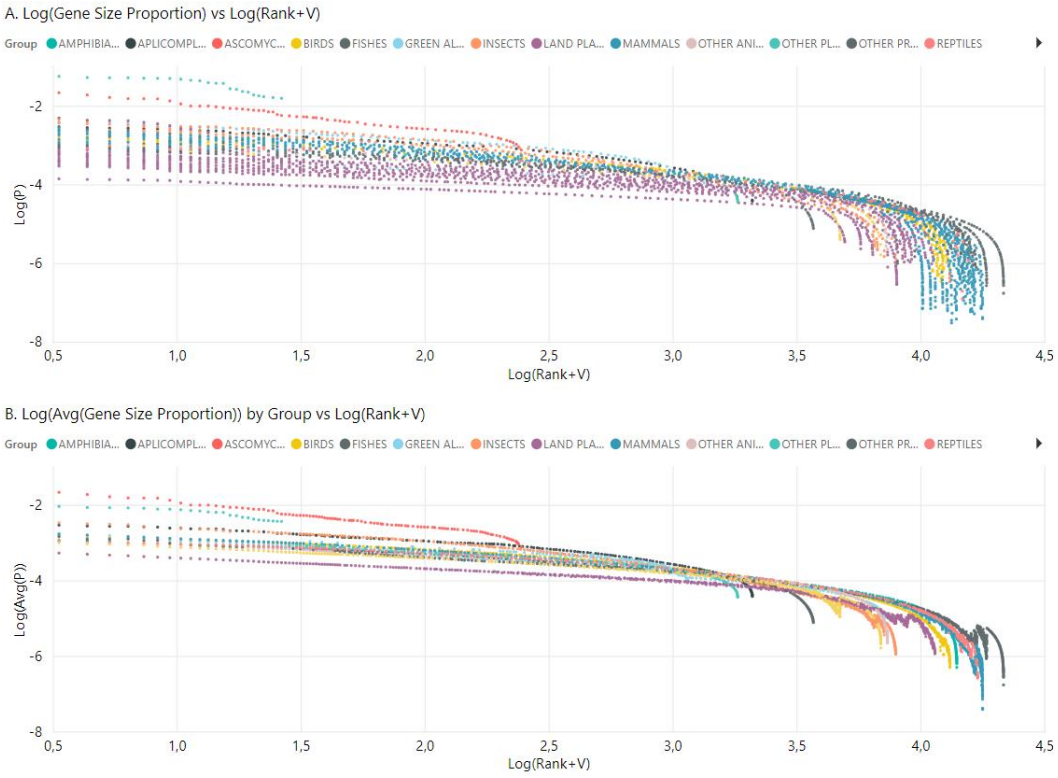
En este trabajo se analizaron las longitudes de todos los genes (enfoque A) y las longitudes de los exones/intrones de los genes interrumpidos (enfoque B) de los organismos, utilizando un abordaje no lineal tipo Z/M, (Figura 6).

Enfoque A

La longitud de los genes de los genomas tiene un comportamiento escalar tipo Z/M

Inicialmente se determinó si la longitud de los genes en un organismo podría tener un comportamiento escalar. Los resultados en la Figura 7 evidencian que las longitudes de los genes en 228 genomas siguen un comportamiento escalar ($R2 \geq 0.75$) y las dimensiones fractales varían entre 0.61 y 1.76. Esto significa que la estructura del conjunto completo de genes de cada organismo tiene una organización estructural no lineal. Teniendo unas curvas más fractales que otras a juzgar, por el valor de la pendiente de la curva y, por ende, por el valor de la dimensión fractal. Esto es, a mayor pendiente de la curva, menor dimensión fractal o grado de escalamiento de los genes en el cromosoma. Así, los grupos de organismos con menor pendiente y mayor D, sus genes tienden a tener tamaños similares, mientras que los grupos de organismos con menor D, tienden a tener genes de tamaños que escalan en magnitud.

Figura 7: Log(Promedio de longitud de los genes) vs. Log(Rank+V) .A) por organismo. B) por promedio por grupo filogenético.

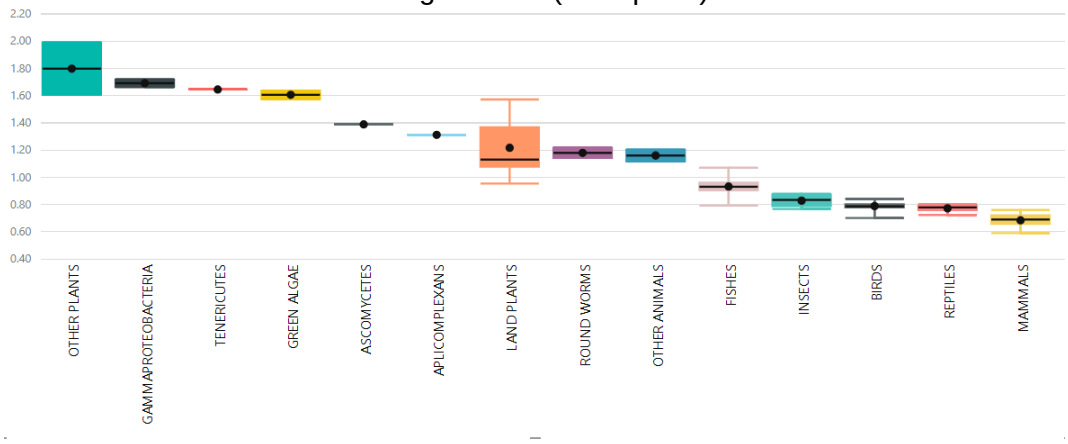


El comportamiento escalar Z/M de la longitud de los genes está de acuerdo con la clasificación filogenética de los organismos

Posteriormente, se examinaron diferentes tipos de organismos a fin de determinar si existe una relación entre la dimensión fractal y la complejidad del organismo. En la Figura 8 se evidencia que el valor promedio de valores de D, por grupo de organismos, exhiben una tendencia inversa en relación con la complejidad del organismo, lo cual es útil para caracterizar la dimensión fractal del grupo de genes en múltiples genomas. Esta distribución se encuentra en concordancia con la clasificación filogenética de las especies (http://tolweb.org/tree). Claramente, las plantas, los hongos, invertebrados y procariotas presentan una ubicación no-fractal, causada por Ds>1, es decir, debido a genes que tiene longitudes similares entre cada taxa. Por el contrario, en los vertebrados (mamíferos, reptiles, aves, peces) e insectos, los genes presentan un comportamiento escalar fractal (D<1) estadísticamente significativo para varios taxa, a juzgar por las cajas y bigotes estadísticos determinadas. Este resultado en principio permite concluir que el análisis de la longitud de los genes aplicando la ley de Z/M es un buen indicador

para el estudio clasificatorio de las especies y las propiedades no lineales implicadas.

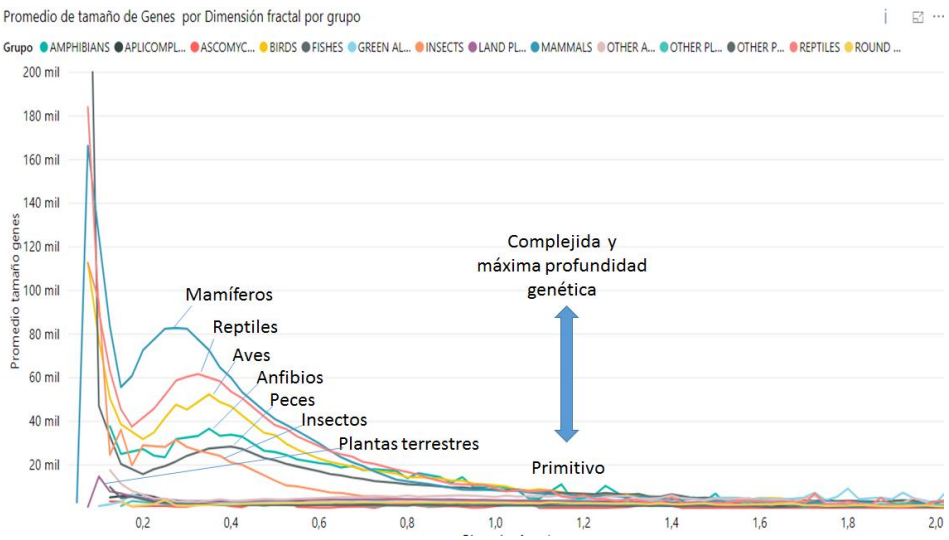
Figura 8: Valores máximo, mínimo y promedio de dimensiones fractales por grupo de organismos (Enfoque A).



El tamaño promedio de los GIs disminuye con la Dimensión fractal - de manera bimodal y asimétrica

La característica diferencial relevante de todos estos análisis está dada por la longitud de los GIs. Por consiguiente, sería interesante averiguar de qué tipo es la relación que existe entre la longitud de los GIs y la D. Para esto, se promedió el tamaño de los GIs por grupo de organismo versus el espectro de D. La Figura 9 muestra una serie de curvas bimodales con dos picos en la región fractal del espectro, con un máximo global (primer pico) y un máximo local (segundo pico) que va disminuyendo en longitud y se va desplazando hacia valores mayores de D en la medida que se desciende en la escala filogenética, lo que muestra que los genes de los organismos más primitivos tienden a ser más cortos que los de aquellos organismos considerados recientes en la escala filogenética. Además, las curvas también muestran un mínimo global y un mínimo local, lo que sugiere el comportamiento de un algoritmo de “optimización por enjambre de partículas”[62] que se discutirá más adelante.

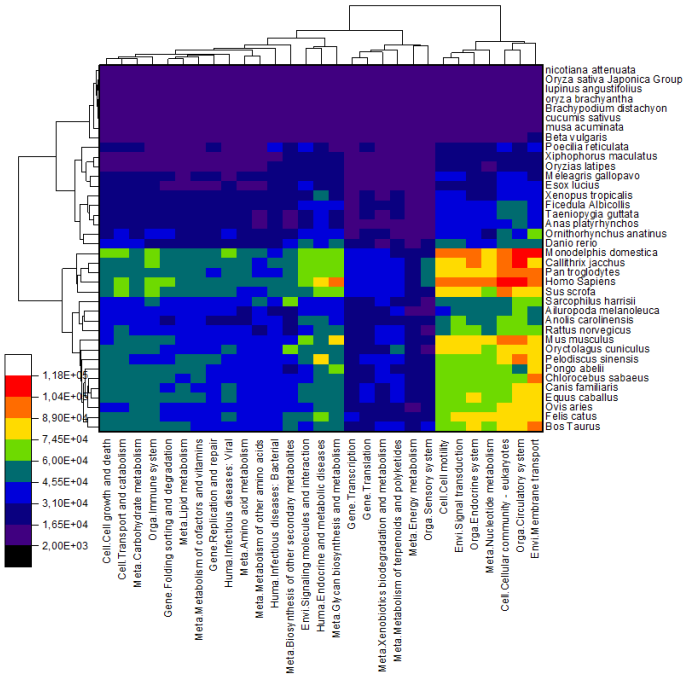
Figura 9: Longitud promedio en pares de bases de genes vs. Dimensión fractal por grupo de organismo.



El tamaño promedio de los GIs disminuye en las Plantas terrestres para todas las rutas metabólicas

Ya que los genes con tamaño pequeño podrían considerarse más primitivos, se averiguo si existe alguna asociación entre el tamaño promedio de los genes y las rutas metabólicas KEGG. La Figura 10 muestra cómo los tamaños se encuentran relacionados con la escala filogenética, siendo los genes más pequeños relacionados con las Plantas terrestres para todas las subcategorías de rutas de KEGG (primitivos). Por el contrario, los genes más grandes se encuentran relacionados con el grupo de humano, chimpancé, jabalí, mono titi y marsupial, entre otros y relacionados con las rutas que determinan la motilidad celular, transducción de señales ambientales, órganos del sistema endocrino, metabolismo de nucleótidos, comunidad celular, órganos del sistema circulatorio y transporte de membrana (ambiental). Otras deducciones similares pueden hacerse para los otros organismos.

Figura 10*:* Mapa de calor con dendrogramas de la Longitud de genes por ruta metabólica versus los organismos. A la derecha se encuentra una guía de la escala filogenética de los organismos utilizados en el mapa de calor.



Organismo	Clase	Nombre común	Orden filogenético
Beta vulgaris	P	Remolacha	1
musa acuminata	P	Plátano	1
cucumis sativus	P	Pepino	1
oryza brachyantha	P	Arroz	1
nicotiana attenuata	P	tabaco	1
lupinus angustifolius	P	Herbácea	1
Brachypodium	P	Graminea	1
Oryza sativa Japonica	P	Arroz	1
Danio rerio	F	Pez	2
Esox lucius	F	Pez-lucio	2
Oryzias latipes	F	Pez-arroz	2
Poecilia reticulata	F	Guppy	2
Xiphophorus maculatus	F	Ciprínido	2
Xenopus tropicalis	A	Rana	3
Anolis carolinensis	R	Lagarto	4
Pelodiscus sinensis	R	Tortuga	4
Anas platyrhynchos	B	Pato	5
Ficedula Albicollis	B	Papamoscas	5
Meleagris gallopavo	B	Pavo	5
Taeniopygia guttata	B	Pinzón	5
Sarcophilus harrisii	M_Mar	D Tasmania	7
Monodelphis	M_Mar	Colicorto	7
Ornithorhynchus	M_Mar	Ornito	7
Bos Taurus	M	Toro	7
Ovis aries	M	Oveja	7
Sus scrofa	M	Jabalí	7
Felis catus	M	gato	7
Homo Sapiens	M	Humano	7
Mus musculus	M	Ratón	7
Pongo abelii	M	Orangutan	7
Equus caballus	M	Caballo	7
Pan troglodytes	M	Chimoancé	7
Canis familiaris	M	Perro	7
Rattus norvegicus	M	Rata	7
Callithrix jacchus	M	Mono tití	7
Chlorocebus sabaeus	M	Mono verde	7
Oryctolagus cuniculus	M	Conejo	7
Ailuropoda	M	Oso panda	7

Las siglas corresponden a las clases filogenéticas: P: Planta, F: Peces, A: Anfibio, R: reptil, B: Aves: M-Mar: mamífero marsupial y M: mamífero. Los números representan el órden en que se encuentran en el árbol filogenético estándar.

Enfoque B

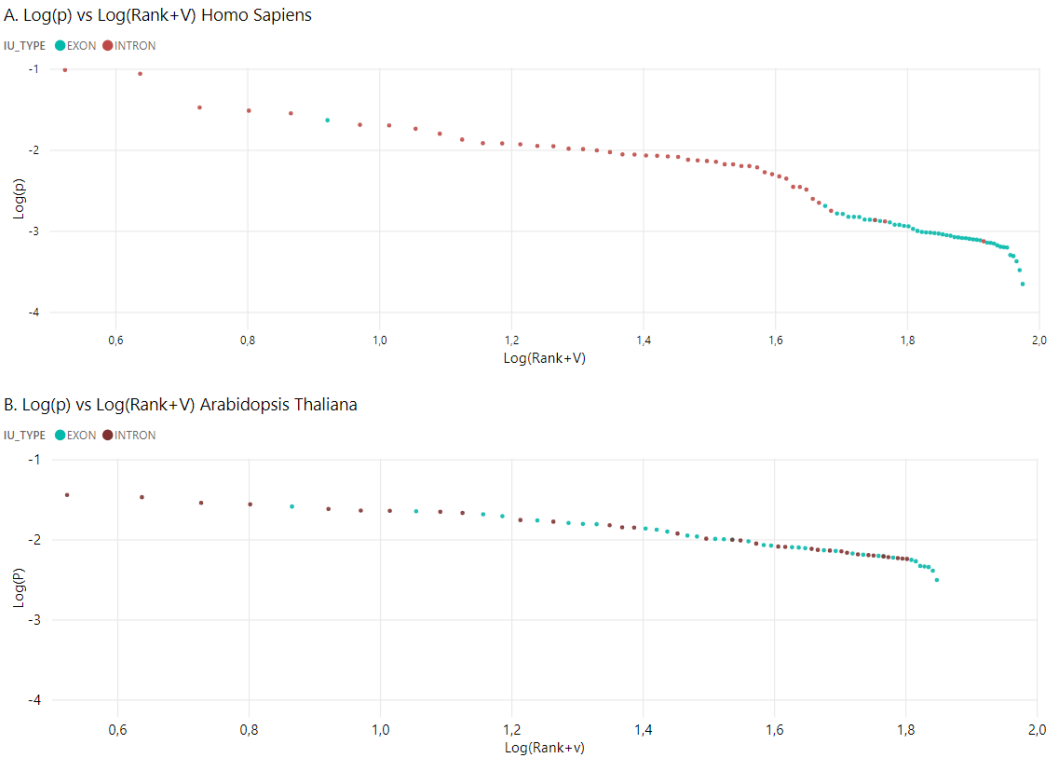
Con el objeto de estudiar de manera sistemática la estructura en longitudes de los exones e intrones de los GIs de múltiples organismos, se procedió a variar los términos de la analogía lingüística (Tabla 2).

La distribución de las longitudes exón/intrón de un gen interrumpido eucariota tiene un comportamiento escalar Z/M

Inicialmente, se procedió a explorar de manera sistemática si dicha distribución no lineal era observada para un simple GI. En la Figura 11 se muestra la distribución de los tamaños de los exones e intrones del gen ABCA1 de *Homo sapiens* y de *Arabidopsis thaliana*. Se encontró que ambos genes presentan un comportamiento bilogarítmico, no obstante, el gen humano es más escalar (en 3 órdenes de magnitud) que el gen de la planta (2 órdenes de magnitud). Se observa una distribución clara del tamaño de los intrones en la parte superior del plano cartesiano y de los exones en la parte inferior (Figura 11A) para el gen humano y

un intercalamiento de los tamaños de intrones y exones para el gen de Arabidopsis (Figura 11B).

Figura 11: Clasificación de la longitud de los genes vs. Clasificación. A. Homo sapiens B. Arabidopsis Thaliana. (Ejemplo de distribución de exones intrones de un gen)

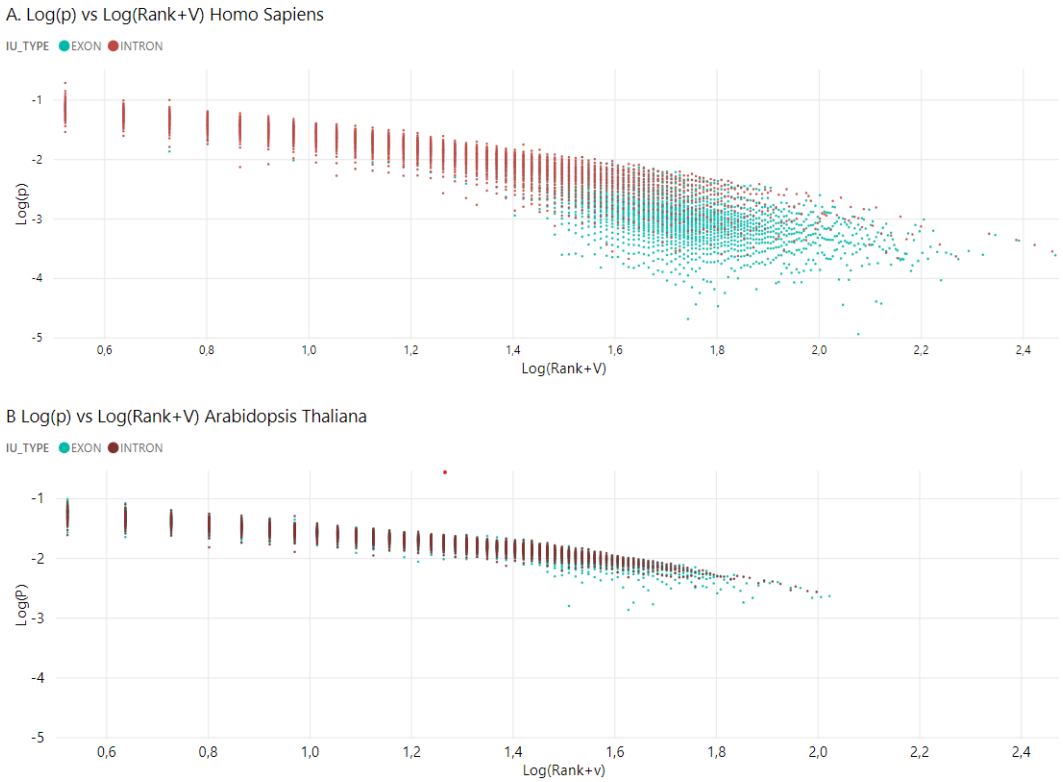


Las longitudes de los intrones-exones de los genes interrumpidos en *Homo sapiens* son más escalares (fractales) que las de los GI de *Arabidopsis thaliana*

Con el objeto de validar si dicho comportamiento escalar observado en un simple gen es extensible a todos los GIs de *Homo sapiens* y de *Arabidopsis* se efectuó un análisis similar con todo el set de genes de cada organismo. La Figura 12 muestra la longitud de los exones e intrones (o unidades de información (UI)) versus el rango de dimensión fractal (D) para *Homo sapiens* y *Arabidopsis thaliana*. Se encontró que las longitudes de las UIs para cada GI de *H. sapiens* tienen una organización fractal, dado que se distribuyen sobre 4 órdenes de magnitud (escalar), donde la gran mayoría de los intrones son graficados en la parte superior de las curvas (las frecuencias más altas), mientras que la mayoría de los

exones son trazados en la parte inferior de las curvas (las frecuencias más bajas), Figura 12A. Por el contrario, el set completo de GIs de Arabidopsis se encuentra distribuido sobre 2 órdenes de magnitud, con un salpicado poco ordenado de exones e intrones, Figura 12B.

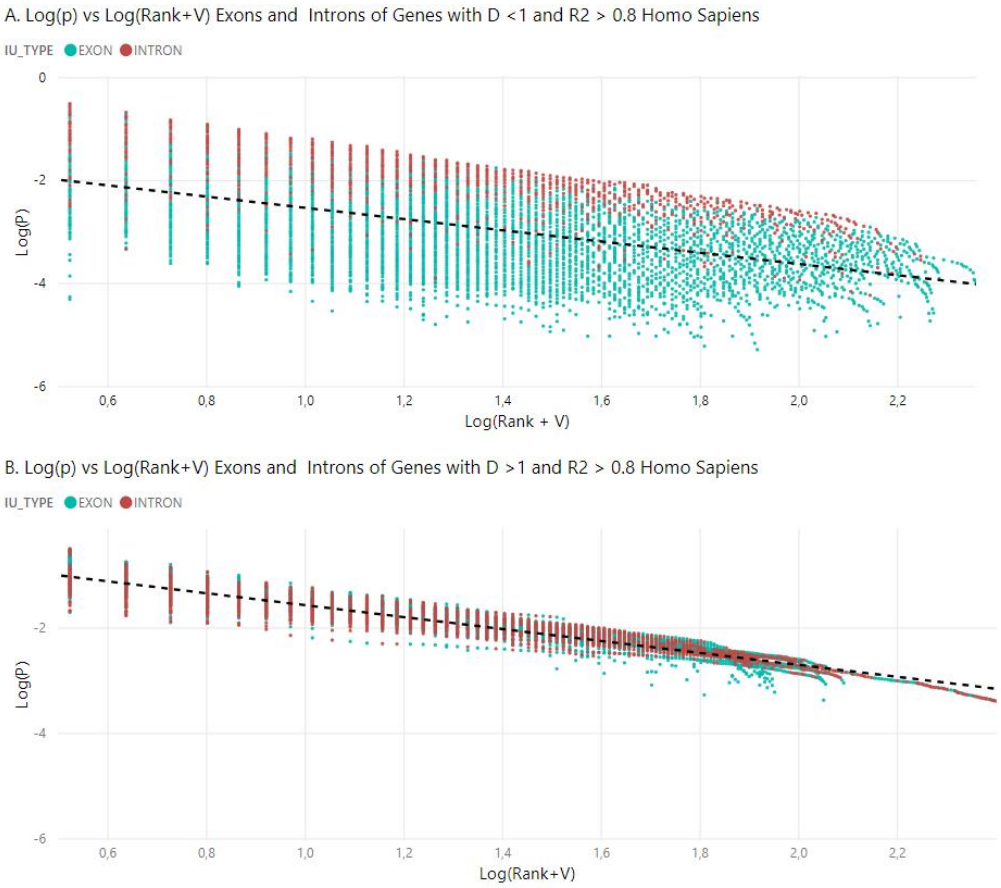
Figura 12: Clasificación de la longitud de los genes vs. Clasificación. A. Homo sapiens B. Arabidopsis Thaliana. (Todos los genes interrumpidos).



La mayoría de los genes interrumpidos de *Homo sapiens* siguen un comportamiento fractal Z/M

Con el fin de examinar que proporción de los GI en Homo sapiens siguen un comportamiento fractal Z/M, se examinaron las Ds para los 20.959 GIs del genoma humano. La Figura 13 muestra la distribución de los GIs humanos con D < 1 (95% -399.175 valores), con una pendiente de -0,46 (Figura 13A) y D > 1 (5% -20.175 valores) con una pendiente de -0.81 (Figura 13B), con un R^2 > 0.8. En conclusión, la mayoría de los GIs de *H. sapiens* (95%) tienen un comportamiento fractal tipo Z/M.

Figura 13: Distribución de los intrones y exones versus la D en H. sapiens para los GIs con D<1 y D>1 y R2 > 0.8. Las líneas de tendencia están indicadas.



La mayoría de las longitudes intrón/exón de los genes interrumpidos de los organismos siguen una ley Z/M tipo fractal

Para examinar si esta distribución escalar es una condición general a los GIs de 219 organismos, se extendió la investigación hasta cerca de 4’232.493 de GIs, Tabla 4. Se encontró que el comportamiento fractal de los GIs depende de los taxones, por ejemplo: en mamíferos el grado de fractalidad es muy alto para el 61% de los GIs (R^2>0.9), mientras que en plantas terrestres es del 48%. Otras consideraciones similares se pueden hacer para los otros grupos de organismos, según el ajuste estadístico. Generalizando, se encontró en toda la muestra de GIs que más de la mitad de todos los GIs (59%) se ajustan al modelo Z/M con R^2 > 0.9, presentando un alto grado de fractalidad (D < 1), un 83% de los GIs tienen un R^2 > 0.8 y un 88% para los GIs con R^2 > 0.7. La cantidad restante de los genes (9.7% con un valor ajustado R^2 > 0.7) es completamente no-fractal (D > 1.0).

Tabla 4: Cantidad total de genes analizados por grupo y rangos de valores D y R2.

Grupo de Organismos		IG D<1 R2>0.9		IG D<1 R2>0.8		IG D<1 R2>0.7		IG D>1 R2>0.7		Total
	Orgs	Cod.Genes	PCT	Cod.Genes	PCT	Cod.Genes	PCT	Cod.Genes	PCT	Genes
GREEN ALGAE	2	6992	0,39	10203	0,57	11024	0,62	6416	0,36	17921
OTHER ANIMALS	2	10455	0,39	16152	0,60	17954	0,66	7831	0,29	27096
ASCOMYCETES	1	132	0,47	175	0,62	217	0,77	13	0,05	283
BIRDS	24	210759	0,59	311094	0,88	326687	0,92	22064	0,06	354557
OTHER PLANTS	2	2624	0,55	3733	0,78	4140	0,87	421	0,09	4756
REPTILES	10	105822	0,60	163861	0,92	172076	0,97	2741	0,02	177322
INSECTS	4	27798	0,59	36661	0,78	39111	0,83	6540	0,14	46973
FISHES	49	696021	0,65	911163	0,85	956249	0,89	92896	0,09	1068731
MAMMALS	91	982397	0,61	1455450	0,90	1530283	0,94	58928	0,04	1619380
ROUND WORMS	2	14908	0,44	21786	0,64	24073	0,71	8269	0,24	33907
APLICOMPLEXANS	1	1290	0,44	1849	0,63	2008	0,68	733	0,25	2939
LAND PLANTS	29	412145	0,48	560530	0,66	611355	0,72	203916	0,24	851163
OTHER PROTISTS	1	5448	0,53	7717	0,75	8613	0,83	1224	0,12	10351
AMPHIBIANS	1	9520	0,56	15116	0,88	16312	0,95	419	0,02	17114
TOTAL	219	2.486.311	0,59	3.515.490	0,83	3.720.102	0,88	412.411	0,10	4.232.493

La primera columna numérica indica el número de organismos analizados, la segunda columna, indica la cantidad de genes que presentan una dimensión fractal D < 1 con un R2 > 0.9, la siguiente columna expresa los datos de manera porcentual (PCT), respecto a la cantidad total de genes codificantes (cod.) analizados. De igual manera, se presentan los datos en las columnas correspondientes para valores de $R^2 > 0.8$ y $R^2 > 0.7$, éste último para D > 1.0 y D <1.0.

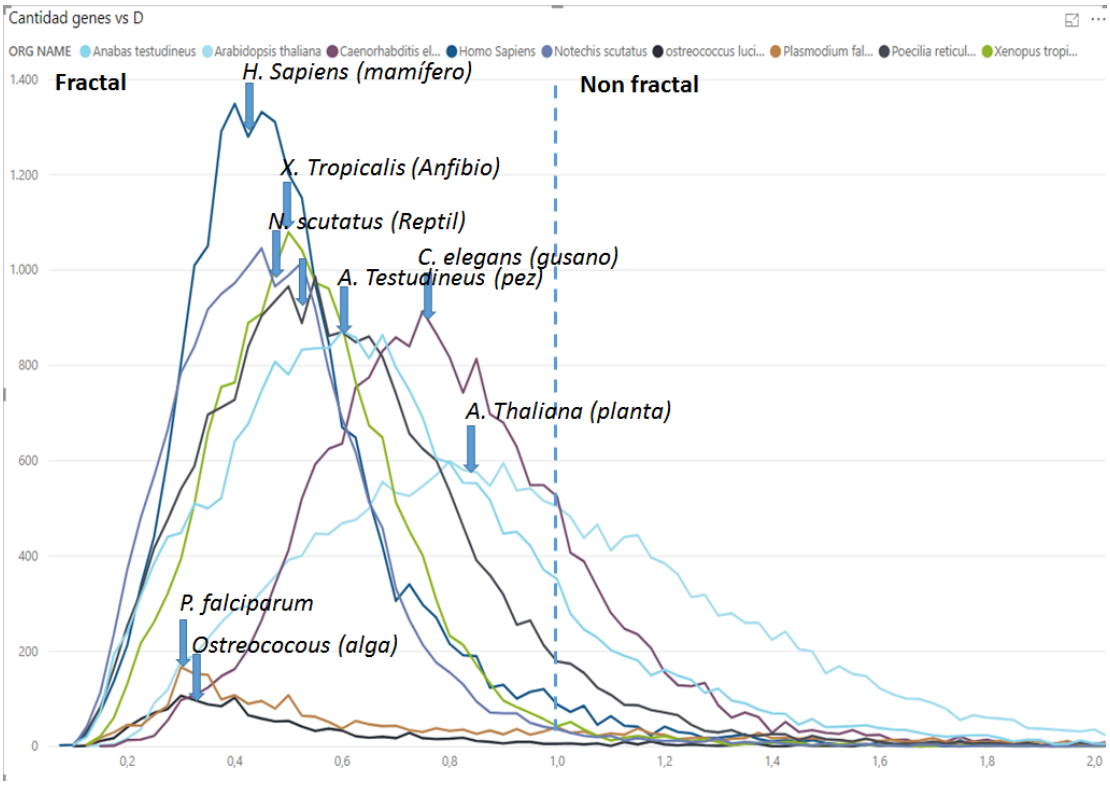
El espectro de los rangos de D revela la existencia de un atractor fractal poblacional para la longitud de los genes interrumpidos

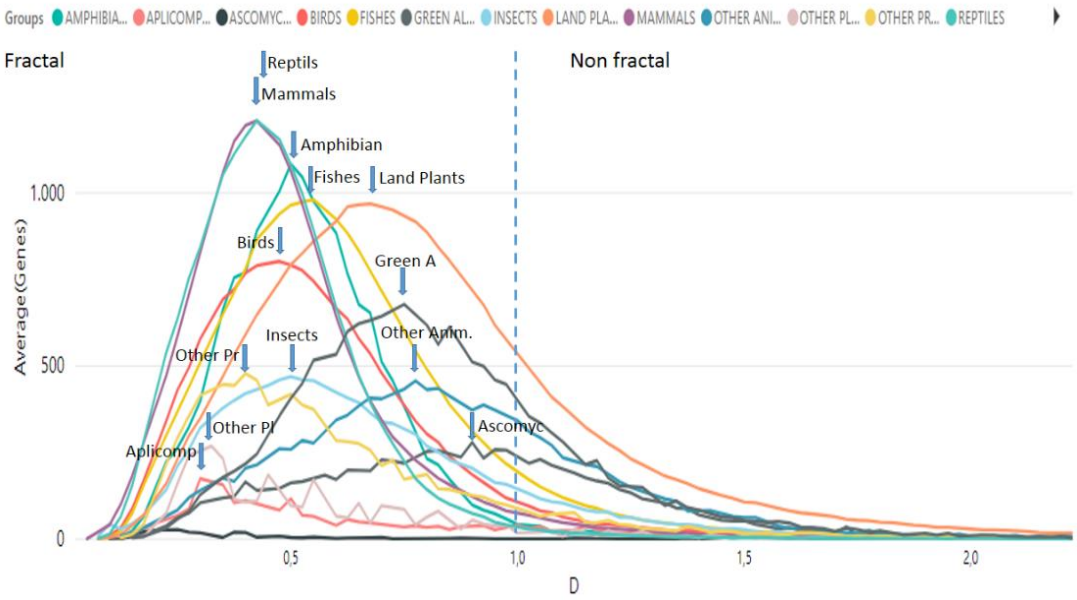
Posteriormente se examinó de manera puntual la distribución del número de genes por rango de D para varios organismos. Se identificó para estos organismos algunos picos de distribución (Figura 14A) lo que sugiere que el espectro de dimensiones fractales podría ser útil para determinar cuántos genes puede haber en un rango específico de valores de D.

Por otra parte, se plantea la hipótesis de si el número promedio de genes por grupo de organismos por rango de D podría estar relacionado con la complejidad

filogenética de los organismos. La Figura 14B permite determinar una serie de curvas poblacionales tipo Poisson para la distribución de los tamaños de los GIs y cuyo valor máximo se encuentra ubicado en la región fractal del espectro de D. Dado que cada grupo de organismo muestra un pico de distribución, a éste se le denominó: “Atractor Fractal Poblacional” (AFP), ya que allí convergen la mayoría del comportamiento escalar de los GIs para cada organismo o grupo de organismos. Es importante recordar que, en un fractal estadístico, el atractor corresponde a las frecuencias más numerosas. Por ejemplo, para los mamíferos los AFPs se encuentran en el rango de valores 0.40 - 0.45, el cual es la organización escalar preferencial presente en la mayoría de sus GIs, para los reptiles de 0.425, para las aves de 0.47, para los anfibios de 0.475, para las plantas terrestres de 0.650, para las algas verdes de 0.78 y para algunos hongos de 0.85, entre otros.

Figura 14: A) Número de GIs vs. Dimensión fractal para algunos organismos (*Arabidopsis thaliana, Caenorhabditis elegans, Notechis sculatus, Ostreococous, Poecilla reticulata, Xenopus tropicalis, Homo sapiens, Anabas Testudineus* y *Plasmodium falciparum*). B) Promedio de genes por rango de D por grupos de organismos.





En todas las curvas de la Figura 14 se observan colas en la región de D>1.0 o región no fractal, lo que hace surgir la pregunta de ¿qué clase de genes podrían estar allí? Pues si existe una relación positiva entre los AFP y la complejidad filogenética de los organismos, como se encontró, es de esperarse que los GIs con D > 1.0 sea genes antiguos o "ancestrales". Para averiguar esto se examinaron a continuación los genes mediante varios enfoques.

Grupos de GIs de rutas metabólicas y funciones celulares se organizan por grupos de organismos

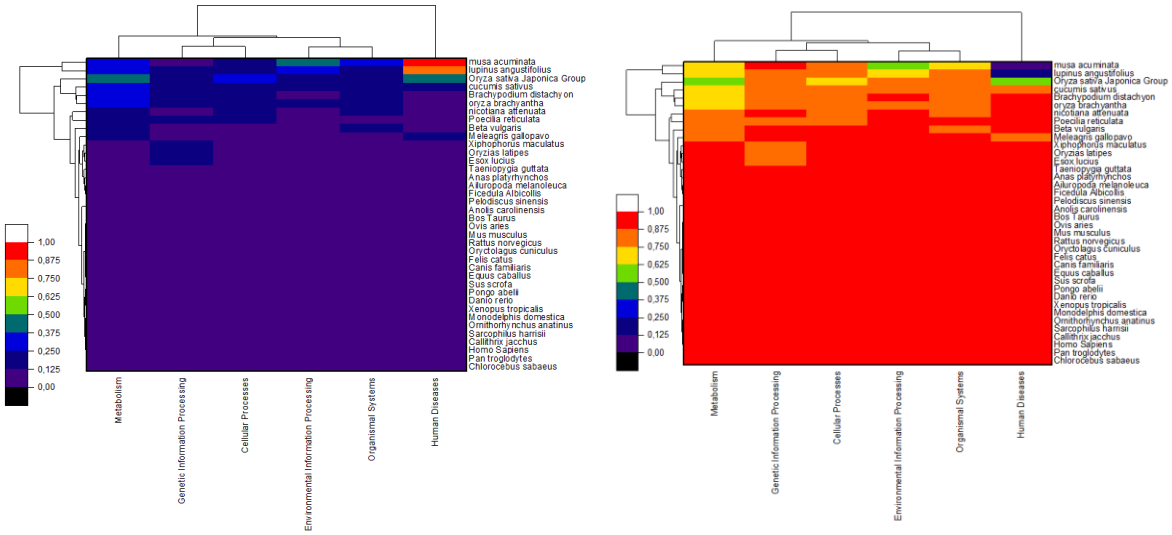
La asociación de rutas celulares a los organismos fue posible mediante la aplicación de un análisis de clúster jerárquico con método de enlace completo, que incluye dos conjuntos de datos en el modelo: i. Las frecuencias de cada GI de cada organismo y ii. Un valor de un gen funcional con base en su asociación con las vías de señalización celular recuperadas de KEGG. Es importante tener en cuenta que este valor fue un puntaje constante con base en la normalización del número de genes por ruta (KO), además se consideró KO como un conjunto de datos seleccionados para un análisis de dimensionalidad. En la Figura 15 y la

se muestran ndos agrupamientos jerárquicos principales para las rutas de señalización de cada celda de GIs con un R^2 > 0.8, que incluyen 6 categorías generales (Figura 15) y 29 subcategorías (

) versus 38 organismos. Este análisis incluye un mapa de calor de frecuencias con un dendrograma de co-asociación.

-Según categorías KEGG

Figura 15: Mapas de calor con dendrogramas del porcentaje de genes por categorías principales (de rutas metabólicas) KEGG versus organismos para A) con D > 1 y para B) con D < 1.



Para D > 1.0 (Figura 15A), las categorías generales presentan 3 subgrupos. Los dos primeros con 0.125 y 0.5 como frecuencias mínimas de configuración (FC) entre los organismos. A pesar del bajo valor, lo más destacado del resultado fue precisamente haber encontrado algún grado de asociación para las categorías "Metabolismo, Procesos celulares, Sistemas de organismos y Procesos ambientales"; donde se muestran ciertos agrupamientos que representan todas las Plantas terrestres (utilizadas en el presente estudio); además, la distribución de los organismos en cada grupo y su frecuencia que define los subgrupos en la mayoría de los organismos parecen seguir la escala filogenética (en un 90%) (Figura 8).

Acerca de la categoría del metabolismo, las vías están principalmente vinculadas con todas las plantas terrestres (FC: de 0.125 a 0.5), lo que podría interpretarse como rutas antiguas y fundamentales para otras actividades [63], [64][65]. Tema que se analizará más adelante.

Finalmente, para el tercer subgrupo, la gran mayoría de los organismos (peces, anfibios, reptiles, aves y mamíferos) se encuentran por debajo de 0.125.

El mapa de calor para D < 1.0 muestra una imagen inversa a la anterior, como era de esperarse (Figura 15B), donde la FC va incrementando de 0.5 a 1.0 desde Plantas terrestres a Mamíferos, de acuerdo con la escala filogenética (en un 90%).

-Según subcategorías KEGG

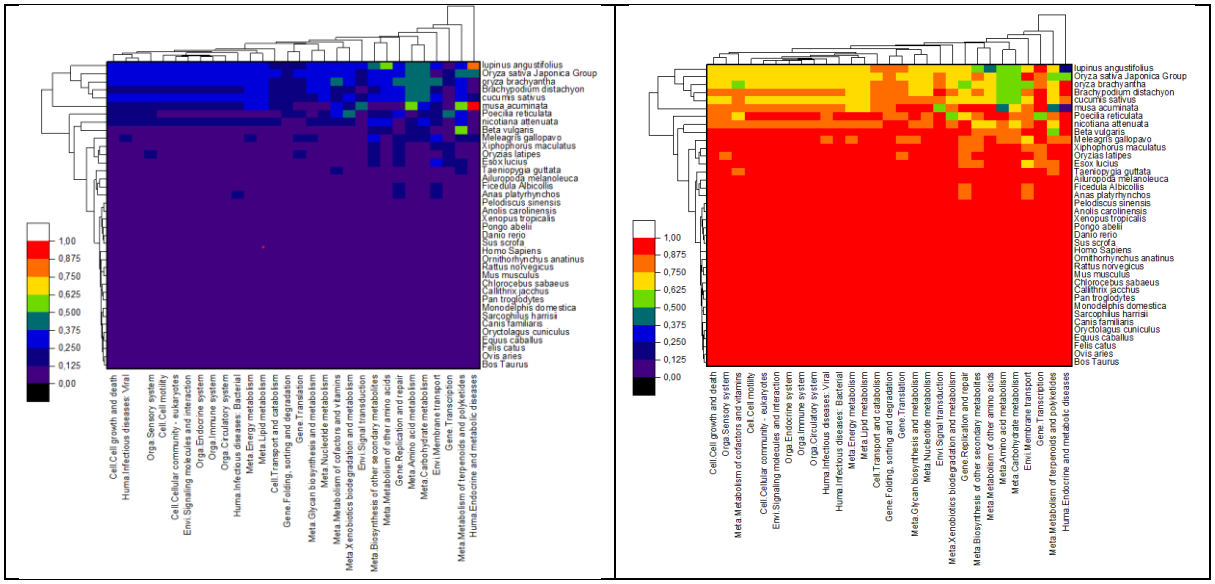
Al igual que los dos resultados anteriores, se examinó la relación entre los GIs de los organismos versus las 29 subcategorías de KEGG. Para D > 1.0 se encontraron FC por encima de 0.25 para la mayoría de las Plantas terrestres y subcategorías, con aumentos puntuales para ciertas rutas y especímenes. Especialmente, para el metabolismo de carbohidratos y de aminoácidos y el metabolismo de terpenoides y poliketides,

A. De nuevo, los GIs de estas rutas metabólicas, podrían ser parte vestigial de genes antiguos.

Para D < 1.0, la

B muestra una imagen inversa a la anterior, donde la FC va incrementando de 0.5 a 1.0 desde Plantas terrestres a Mamíferos, de acuerdo con la escala filogenética (en un 90%).

Figura 16: Mapas de calor con dendrograma del porcentaje de genes por subcategorías principales (de rutas metabólicas) KEGG versus organismos para A) con D > 1 y para B) con D < 1.



Grupos promedios de dimensiones fractales de rutas metabólicas y funciones celulares se organizan por grupos de organismos

La asociación de rutas celulares a los organismos se hizo mediante la aplicación de un análisis de clúster jerárquico que incluye un modelo simple: Se tomaron los GIs que pertenecen a la misma ruta metabólica por organismo y se sacó el promedio de D. Aquí se observa en términos de promedios de D la misma tendencia descrita anteriormente, donde los organismos más complejos tienen como promedio un valor de D más bajo, el cual se encuentra de acuerdo con la clasificación filogenética de las especies, donde las Plantas terrestres tienen las categorías (

Figura **17**A) y subcategorías (

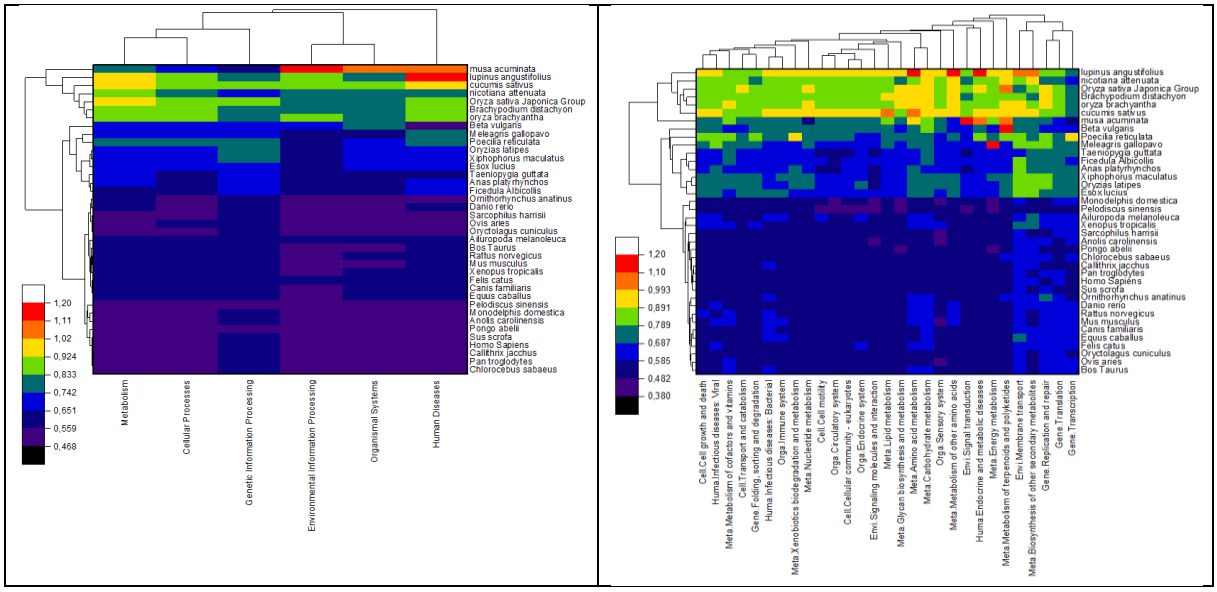
Figura **17**B) KEGG, con la Ds más altas. Es relevante notar que los valores promedios de D superior a 0.742 (

Figura **17**A) y promedios de D superiores a 0.789 (

Figura **17**B) pertenecen a las Plantas terrestres. Entre estos se destacan los GIs relacionados con el Metabolismo y los Sistemas de organismos en las categorías y el Metabolismo de lípidos, de aminoácidos, de carbohidratos, de energía, terpenoides y poliketes, transporte de membranas, replicación y traducción en las subcategorías. Especial relevancia tienen aquellos GIs con promedios de D > 1.0, lo cual los convierte en candidatos para ser GIs muy “antiguos o ancestrales”,

entre los cuales tenemos: transporte de membrana, metabolitos secundarios, metabolismo de aminoácidos, metabolismo de lípidos, terpenoides y poliketides.

Figura 17: Mapa de calor con dendrograma del promedio de D por organismo y ruta metabólica KEGG. A) Por categorías y B) Por subcategorías.

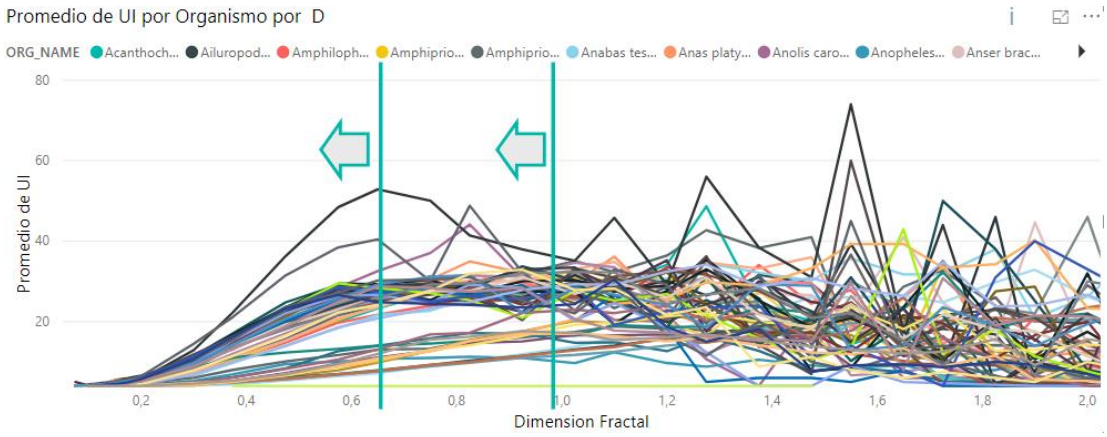


El promedio del número de las UIs está directamente relacionado con el valor de D

Sin duda la organización modular de los GIs en exones e intrones es un problema de fragmentación de la información genética. Los GIs podrían colectivamente considerarse formados por unidades de información (UI). Por otra parte, llama la atención que la ley de Z/M aplicada a los GIs (enfoque B) se cumple para la gran mayoría de los GIs (88%) y no se cumple por lo general para una minoría (~10%) de los GIs (Tabla 4). Con el objeto de estudiar el problema de la fragmentación del GI se determinó el número promedio de UI versus la dimensión fractal. Se encontró de manera sorprendente que dicha relación en la región fractal sigue una relación lineal directa para todos los GIs, Figura 18, lo que significa que al aumentar la D aumenta el grado de fragmentación del gen interrumpido. Esto es particularmente cierto entre los rangos de $0<D<0.65$ y un poco por debajo de $D=1$. Además, dicha fragmentación aumenta con la complejidad de la escala filogenética, es decir, los organismos más simples tienen menor número de UI en los GIs, mientras que los organismos más complejos tienen un mayor grado de fragmentación de los GIs. Es notable observar que por encima de $D>1$ el grado de

fragmentación de los GIs pierde su correlación en la región no fractal para la gran mayoría de los organismos, excepto (y en parte) para algunos organismos considerados más primitivos.

Figura 18: Promedio de unidades de información por rango de D por grupo de organismos.



2.4 DISCUSIÓN Y CONCLUSIONES

En este trabajo se utilizaron dos analogías entre la estructura del ADN y un lenguaje. El enfoque A utilizó la longitud total del gen para hallar la dimensión fractal del organismo y el enfoque B utilizó la longitud de los intrones y exones para hallar la dimensión fractal de los GIs. Ambos enfoques mostraron resultados coherentes confirmando la validez de las analogías.

Enfoque A

Inicialmente, se descubrió que la longitud de los genes interrumpidos de los eucariotas, siguen un comportamiento escalar Z/M, donde las frecuencias más altas estarían dadas por los genes más grandes, mientras que las frecuencias más bajas por los genes más pequeños. Dicho comportamiento está de acuerdo con la clasificación filogenética de los organismos, donde los genes más grandes se encuentran en los organismos más recientes, como mamíferos, reptiles, aves, insectos y peces y los genes más pequeños, en los gusanos redondos, las plantas terrestres, hongos, protistas, algas y bacterias (Figura 8). Este resultado que se encuentra de acuerdo con lo esperado ya que estos últimos organismos se consideran más primitivos que los animales superiores.

Según la analogía lingüística de Mandelbrot citada en la metodología, los textos con D>1.0 serían textos no fractales, finitos y contaminados o mutados, por ende, textos pobres en información, debido a que las frecuencias de las longitudes (de las palabras o genes) tienden a ser similares. Así, cuando la estructura modular en la longitud (de la palabra o) del gen es similar (muy ordenada), es difícil construir diversidad biológica en términos filogenéticos (o riqueza de vocabulario). Hubo entonces la necesidad que la longitud del gen escalara en varios órdenes de magnitud (hacia la región fractal o "desordenada") para poder enjambrar o dar explosión a una diversidad de estructuras modulares de genes y proteínas y aumentar así la riqueza del repertorio del "vocabulario molecular del ADN". En consecuencia, es notable que en la región no fractal se encuentren los genes de los organismos considerados más primitivos, mientras que en la región fractal se encuentren los animales más recientes y la diversidad biológica más alta.

- **¿Cómo pudo haber surgido una diversidad de estructuras de GIs tan variada?**

Continuando con la analogía lingüística de Mandelbrot, los textos con D<1.0 serían textos **fractales e infinitos**, con un grado de complejidad que oscila entre 0 y 1, donde los textos más complejos en términos lingüísticos tienden a 1.0, mientras que los textos óptimos, regulares y de máxima organización tienden hacia la parte media de la escala (0.4 – 0.6) y los extremadamente simples, por debajo de 0.4 [19], [66]. En el caso de los GIs de los organismos, el "corrimiento" de D, desde D = 1.0 hacia 0 comenzaría con un lenguaje infinito de estructuras las cuales se fueron "materializando" en peces, insectos, anfibios, reptiles y aves hasta alcanzar la parte media del espectro de D a los mamíferos y al *Homo sapiens* Figura 8.

- **¿Cómo pudo haberse dado un cambio del lenguaje estructural del gen de lo no fractal a lo fractal?**

Mandelbrot establece que los discursos con valor de D = 1.0 o muy cercanos a 1 se encuentran en el punto crítico o transición de fase de donde se pasa de lo no fractal a lo fractal [67][68]. En ensayos de simulación de dinámica de agregación se han generado reacciones que atraviesa el punto crítico y se puede mover en ambas direcciones [69], lo que representa un significativo paso unificador hacia una teoría de escala completa del crecimiento fractal y la formación de patrones lejos del equilibrio. Similar a esto, la transición de fase cercana a D = 1.0 pudo haber llevado a la complejización (crecimiento) de los genes desde lo no fractal a

lo fractal, generando “semillas” [70] de secuencias de ADN autónomas que se fueron replicando y escalando en longitud. En las reacciones químicas de estructuras disipativas (o en transiciones morfológicas simuladas por dinámica de agregación), una vez dadas ciertas condiciones iniciales conducidas por transiciones fractal/no fractal (desordenadas/ordenadas), propician que la reacción de agregación (para el caso, longitudes de genes) se desplace en una u otra dirección [69].

- **¿Cómo pudieron aparecer estructuras ordenadas?**

Otra propiedad de las estructuras disipativas es su extensión de la teoría termodinámica a sistemas alejados del equilibrio, que sólo pueden existir en conjunto con su entorno [71][72]. Esta propiedad se podría extender a la estructura del gen, a fin de promover la aparición de estructuras coherentes autoorganizadas como sistemas dinámicos no lineales que se alejan del equilibrio[72], en esta caso, llamados genes. Así, la disipación de la energía y la materia que suele asociarse a la noción de pérdida y evolución hacia el desorden se convierte, lejos del equilibrio, en fuente de orden [72]. Como se mencionó, en una disolución la reacción química oscila rítmicamente entre dos extremos diferentes. En una película plana emergen patrones cambiantes, ondas y estructuras espaciotemporales. Estos patrones son muy sensibles a las condiciones iniciales, donde una pequeña perturbación localizada genera un gran cambio global en el patrón observado. Así, la teoría del caos estudia la generación de patrones, oscilaciones, auto semejanzas y correlaciones a larga distancia que se encuentran en el orden y el desorden. Los sistemas caóticos son deterministas pero difíciles de predecir ya que presentan sensibilidad a las condiciones iniciales. Muy seguramente, los ácidos nucleicos pudieron haber experimentado este tipo de transiciones deterministas. Vale recordar que los ácidos nucleicos son polímeros capaces de haber sido polimerizados por la corteza terrestre [73]. De esta manera, pudo suceder que la geoquímica de la Tierra –en el pasado- favoreció fuertemente que las reacciones de polimerización de los ácidos nucleicos (y las proteínas) se dieran en la dirección de lo no fractal a lo fractal, una vez alcanzado el punto crítico. Esto pudo haber sucedido durante la explosión del precámbrico y el cámbrico que llevaron a la aparición repentina, rápida y diversa de organismos macroscópicos multicelulares complejos, hace 542/530 millones de años [74][75]. Este periodo marca una brusca transición en el registro fósil y geológico con la aparición de los miembros más primitivos de los animales multicelulares (metazoos)[76]. El algoritmo PSO* por excelencia es aquel aplicado a las abejas en busca de alimento, el cual podríamos extenderlo a los genes (las abejas o las

partículas) que se fueron localizando en entornos de espacios definidos (buscar alimento), hasta conformar un colectivo de enjambre (de genes) en una región del espacio explorado, hasta encontrar una región (el celular), con óptimo local conocido por el enjambre (óptimo global) que lo orientará hacia una nueva dirección, hasta encontrar su máxima complejidad autoorganizada. El valor máximo local parecería estar también relacionado con la complejidad del organismo o máxima profundidad posible en el contenido de información algorítmica (según M. Hellman, 1995) o genómica, para el caso que nos ocupa, ya que su valor aumenta gradualmente siguiendo la escala filogenética o disminuye según la antigüedad del taxa (primitivo). “De esta manera, estos procesos podrían haber generado caminos con memoria molecular que llevarían los genes a establecer especies únicas de organismos”. Es decir, el caos determinista.

- **¿Cómo funcionaría esto?**

Supongamos que tenemos un mecano (un GI) que tiene tres parámetros que rigen su funcionamiento (tamaño promedio del GI (masa), dimensión fractal (energía) y otra propiedad, por decir algunas): Gracias al PSO podemos calcular cuáles son los valores de esos parámetros para que el sistema tenga un consumo mínimo. Este algoritmo tiene modificaciones que hacen que los mecanos "aprendan" o que los mecanos no se atasquen en ciertas posiciones del problema a resolver: “establecer la celularidad” (encontrar alimento). La fuerza con que los GIs (partículas) son empujadas en cada una de estas direcciones (de la curva bimodal) depende de dos parámetros que pueden ajustarse (**atracción-al-mejor-local** y **atracción-al-mejor-global**), de forma que a medida que los GIs se alejan de estas localizaciones mejores, la fuerza de atracción es mayor[12]. Este mecanismo podría ser favorecido si el atractor global y local son de naturaleza fractal, como de hecho se deduce de la Figura 9.

- **¿Y qué hay de las rutas metabólicas?**

Según las Figuras 9 y 10 las rutas metabólicas varían de acuerdo con la escala filogenética. Con respecto a las rutas metabólicas KEGG sólo se pudo trabajar la información a partir de las plantas terrestres hasta mamíferos. No fue posible de momento, utilizar las bases de datos de organismos considerados muchos más primitivos, lo cual hubiera dado una mayor definición al estudio. Así que la discusión al respecto sólo se puede plantear a partir de las plantas, la cual está de

[12] www.cs.us.es/ ~fsancho/?e=70

acuerdo *grosso modo* con la escala filogenética estándar. La figura 10 lo que dice es que todas las subcategorías de KEGG se encuentran representadas en todos los organismos analizados, no obstante, muestra que en todas las rutas metabólicas hay genes que difieren en longitud para satisfacer todas las estructuras y funciones del metabolismo analizadas. Esto podría interpretarse, como que las bioquímicas de los organismos hicieron sus reajustes (o adaptaciones) para satisfacer las necesidades básicas impuesta por su entorno. De hecho, un alineamiento de secuencias de un mismo gen podría fácilmente revelar que las longitudes de los genes (o las regiones conservadas) van cambiando según la escala filogenética, conservando cierta ortología particular (homología) y reajustado (su longitud) a cada especie en particular.

Enfoque B

Inicialmente, se descubrió que las longitudes de los intrones y exones de un simple GI sigue igualmente un comportamiento escalar Z/M (Figura 11). Luego esta escalaridad fue encontrada a nivel de casi todo el set de genes de un organismo (Figura 12 y Figura 13) y finalmente de toda la muestra analizada (Tabla 4).

- **Los exones e intrones de los GI pueden interpretarse con analogías lingüísticas**

Estos hallazgos tienen profundas implicaciones cuando son interpretados en términos de las analogías lingüísticas expresadas arriba. Según Mandelbrot, un discurso o texto se puede caracterizar por la presencia de 2 componentes relacionados. i) **Conectividad y gramaticalidad**: estos significan que las palabras con frecuencias más altas implican la conectividad del habla y las frecuencias más bajas involucran la gramaticalidad del texto. De acuerdo con la biología molecular conocida de los GIs en eucariotas, el papel de los intrones (son silentes) es conectar los exones, mientras que el papel de los exones es llevar la información codificante para el ARNm que será traducido a proteína. Por ejemplo, en vertebrados, la longitud de los intrones es generalmente mayor a la longitud de los exones (Figura 12)[59], como resultado, la conectividad y gramaticalidad están determinadas preferentemente por los intrones y exones, respectivamente; tendencias que pueden variar en otros organismos más primitivos (Figura 13B) que tienen un comportamiento menos escalar. Por lo tanto, la escalaridad, conectividad y gramaticalidad serán dependientes de los taxones. De hecho, es

conocida la distribución estadística de la longitud de exones e intrones de muchas especies de organismos [77]; [58]. Estas variaciones en la arquitectura de los GIs pueden obedecer a requerimientos posicionales de diferentes aparatos de splicing nuclear, a un reflejo de la historia y regulación génica y/o a su contenido de ADN repetido, como ha sido reportado previamente[59]. Por otra parte, es conocido que aquellos GIs que presentan una alta proporción de intrones de longitud corta y exones de longitud larga, podrían estar relacionados con genes altamente expresados[60], pero estos generalmente no son la norma.

Otro aspecto que plantea Mandelbrot para evaluar un discurso es ii) **Determinismo y variación**: esto es, que, en un lenguaje de comunicación humana, las frecuencias más altas implican determinismo y las frecuencias más bajas involucran variación, tal como se observa en artículos, preposiciones y palabras, respectivamente. En el caso de los GIs, los intrones (secuencias invariantes específicas para cada especie) podrían involucrar cierto grado de determinismo, mientras que los exones podrían ser responsables por la variación genética del mensaje genético. De hecho, lo son. Esta afirmación es correcta en aquellos organismos donde los intrones son generalmente más largos que los exones, como en los vertebrados, pero podría tener un menor grado de relación en plantas y algunos invertebrados, donde los exones pudieran tener las mayores longitudes.

- **Los genes con un valor de D > 1.0 podrían ser ancestrales**

Según Mandelbrot, los lenguajes con un valor de D > 1.0 son lenguajes contaminados, como aquellos que contienen palabras con diferentes significados [19]. Debido a que no existe escalaridad con D > 1.0, se consideró este hecho como un rasgo primitivo en la arquitectura de un GI. Estos genes “ancestrales” se observaron en todos los organismos, pero muy especialmente en plantas, gusanos, algunos insectos y protistas. Estos genes generalmente codifican proteínas relacionadas con: Metabolismo de lípidos, de aminoácidos, de carbohidratos, de energía, terpenoides y poliketes, transporte de membranas, replicación y traducción y metabolitos secundarios. Por ejemplo, los terpenoides y poliquetos son genes muy relevantes en la aparición de las angiospermas (plantas con flores)[74]. Estas rutas de moléculas y enzimas están involucradas en una bioquímica más primitiva que aquellos genes con valores de D < 1.0, involucrados en procesos más especializados. Por ejemplo, cerca del 50% de los GIs en Arabidopsis tienen valores de D > 1.0, sugiriendo una bioquímica

altamente óxido-reductiva, debido a la presencia de una gran cantidad de enzimas relacionadas con el proceso de la fotosíntesis[76].

Como se mencionó anteriormente, Mandelbrot [19] establece que aquellos lenguajes con valores de D < 1.0 son infinitos, mientras que aquellos genes con valores de D ≥ 1.0 son finitos. El descubrimiento de GIs con valores de D mayores o menores que 1.0 y el "movimiento" de los AFPs a lo largo del espectro de complejidad de D lleva a proponer que los GIs se complejizaron desde regiones no-fractales hacia regiones fractales, seguidos por plantas, invertebrados y vertebrados. Según este escenario hipotético, aquellos GIs con valores de D > 1.0 corresponderían a GIs (exones e intrones) ancestrales o fosilizados, en comparación con aquellos GIs que presentan estructura fractal. De esta manera, el genoma "aprende a hablar" empleando un lenguaje infinito (o fractal) por fragmentación escalar para codificar la mayor diversidad biológica posible. No obstante, excepciones a esta regla podría ser los casos de la levadura, algas y protistas, debido a sus roles de parásitos unicelulares e intracelulares, de algunos de ellos.

- **El debate del origen y evolución del intron sigue abierto**

Es importante discutir desde esta perspectiva el origen dinámico de la fragmentación de los GIs, o dicho en otros términos por W. Gilbert: "why genes in pieces?". El origen de los intrones se debate entre las hipótesis de intrones tempranos (introns-early) e intrones tardíos (introns-late) [78]. Estudios recientes se inclinan a favor de la hipótesis de intrones tempranos [60]. Por ejemplo, se han logrado algunos progresos para estudiar la arquitectura del gene eukariota, mediante la combinación del análisis comparativo de numerosos y diversos genomas de eucariotas, reconstrucciones probabilísticas de ganancias y pérdidas de intrones y la teoría genética de la evolución de la complejidad genómica de la población no adaptativa [79]. Los autores llegan a la conclusión de que la evolución de los eucariotas en su conjunto, así como la evolución de cada uno de los supergrupos eucariotas, comenzó con genomas ricos en intrones con señales de empalme relativamente débiles y propensas a errores. Los autores se preguntan: "¿Cuáles son los mecanismos de pérdida y ganancia de intrones? Y consideran que hay muy poca evidencia directa alguna. Parece existir un consenso con respecto al papel de la transcripción inversa en la pérdida de intrones, aunque incluso este mecanismo necesita corroboración experimental. En cuanto a los mecanismos de ganancia intrónica, a pesar de las indicaciones de la

participación de la reparación de rotura de doble cadena, el estudio de este problema clave ni siquiera ha comenzado en serio. La elucidación del escenario general de evolución de la arquitectura del gen eucariota de ninguna manera implica que los principales problemas en el estudio de la evolución y la función del intrón se han resuelto. Por el contrario, las preguntas fundamentales siguen abiertas".

Queda claro que se necesitan novedosos abordajes para atacar el problema del origen de la arquitectura del gen interrumpido. El presente trabajo abordado desde una perspectiva no lineal, no se inclina por ninguna de las dos hipótesis mencionadas arriba. Por el contrario, lo que se podría plantear aquí es que tanto los exones como los intrones parecen haber emergido simultáneamente en organismos primitivos y con dimensiones similares y extremadamente cortas, a fin de satisfacer la función del intrón en generar la estructura lariant (o de lazo) durante el procesamiento del ARNhn (o tipos de splicing), la protección del exón de las mutaciones [80]y posibilitar el transporte del ARNm desde el núcleo al citoplasma para la síntesis de proteínas.

- **¿Cómo intentar visualizar una mecánica subyacente no lineal entre enfoque A y el enfoque B?**

Es decir, ¿cómo los exones y los intrones se integran para generar los genes interrumpidos? Tomando en consideración todos los elementos discutidos anteriormente se podría proponer el siguiente escenario. Las transiciones de fases de las estructuras disipativas es un primer requerimiento para "lanzar" los exones e intrones (desde una región no fractal) a aumentar sus longitudes (por agregación isotrópica/anisotrópica) y "ensamblarse" de manera fractal en GIs (más grandes), mediante un "mecanismo de atractores PSO" que luego serían anidados en celularidades o semillas generadoras de organismos de especies específicas. Esto, debido a que la propiedad fractal posibilita la existencia de un lenguaje infinito de posibilidades (de palabras código predefinidas) para generar de manera determinista (caos autoorganizado) la información que codifica los organismos. Así, en el remoto pasado, la vida pudo haber sido auto ensamblada (mediante palabras código predefinidas) y enjambrada in situ o sembrada (como semillas) en la Tierra para dar origen a cada una de las especies existentes. Todo aquí podría ser especulativo, no obstante, este escenario tiene la bondad de ser derivado de los resultados del estudio y su interpretación hecha a la luz de las propiedades no lineales observadas en otros estudios, los cuales dan luces para plantear nuevas

hipótesis y escenarios. Con este trabajo se propende establecer una teoría hacia la fragmentación e integración no lineal de la información genética (del GI) de camino hacia la celularidad y la especie, propuesta que sin duda entra en controversia con la teoría de la evolución de las especies (por mutaciones aleatorias, dirigidas por selección natural y aislamiento geográfico reproductivo), debido al componente determinista que entraña la teoría del caos.

Finalmente, existen otros aspectos a tener en cuenta, que hacen parte del rompecabezas y no se abordaron en la presente investigación, las **palabras código predefinidas** [81], es decir, cómo las UI de exones e intrones y motivos regulares (en el ADN y la proteína) emergieron para dar cuenta de toda la actividad metabólica (interactómica) que caracteriza la célula viva. En otras palabras, cómo desencriptar el programa de programas que establece el diccionario de motivos y la regulación genética de la información para la vida. Ya, el proyecto ENCODE[13] y [82], entre otros autores, han descifrado gran parte de los elementos funcionales y estructurales codificados por el genoma humano. No obstante, este tipo de estudios acaba de comenzar y se deben de extender a otras especies a fin de establecer una generalización de las palabras código predefinidas para todas las especies.

Conclusiones

Se examinó la aplicabilidad de la ley de Z/M en el problema de la longitud del GI, para cuantificar el grado de fragmentación de este en intrones y exones.

Se implementaron dos enfoques lingüísticos aplicando la ley de Z/M. El enfoque A reveló una relación positiva entre los valores de D y la complejidad de los 228 organismos. Es de resaltar que la distribución escalar de los GIs se encuentra totalmente en concordancia con la clasificación conocida de las especies.

El enfoque B permitió efectuar la medición del grado de fragmentación (en intrones y exones) de millones de GIs, lo que significa que el ARNm y las proteínas parecen estar formados por partes o módulos de información genética escalables. Sería interesante averiguar si existe una relación entre dichos módulos y la estructura de las proteínas, tal como W. Gilbert postuló tiempo atrás.

[13] CODIGO: Enciclopedia de elementos de ADN: www.encodeproject.org/

Adicionalmente, la utilización de un espectro de rangos de valores de D permitió identificar *Atractores Fractales Poblacionales* (AFP) para cada distribución de genes por organismo, así como examinar diversas hipótesis relacionadas con la cantidad promedio de las UIs y la longitud promedio de los genes discutidas en términos de la teoría lingüística de Zipf/Mandelbrot.

Es importante señalar la existencia de GIs no-fractales, los cuales sugieren una arquitectura primitiva relacionada con una bioquímica de vida más simple. Estos resultados sugieren que el lenguaje estructural de la vida, dado por la arquitectura modular de los GIs se complejizó desde regiones no-fractales (en plantas e invertebrados) hacia regiones fractales (en vertebrados), coincidiendo con la clasificación filogenética de las especies. De esta manera, el genoma "aprende a hablar" por fragmentación escalar de la "palabra", hasta alcanzar su máxima complejidad en los mamíferos.

Todo esto da indicio de la existencia de una estructura subyacente de un lenguaje estructural que determina la arquitectura de los genes interrumpidos.
Finalmente, se hizo uso de las propiedades de los sistemas dinámicos no lineales, derivadas de las teorías de la complejidad, la teoría del caos, la geometría fractal, los atractores, las estructuras disipativas y autoorganizadas y la lingüística para interpretar los resultados y generar nuevos marcos de referencia e investigación.

3 CARACTERIZACIÓN MULTIFRACTAL DE POBLACIONES HUMANAS

3.1 INTRODUCCIÓN

La caracterización de poblaciones humanas ha sido abordada previamente desde diferentes enfoques, como los estudios de asociación del genoma completo (en inglés, **GWAS** (Genome-wide association study) o WGAS (Whole genome association study) [83], los cuales han revelado múltiples asociaciones de variantes genéticas con enfermedades o estados saludables. Sin embargo, el estudio más representativo hasta el momento se ha enmarcado en el proyecto 1000 Genomas. Dicho proyecto buscaba proporcionar una descripción de las variaciones genéticas humanas más comunes, secuenciando completamente 2.504 individuos pertenecientes a 26 poblaciones diferentes. Como resultado

encontraron más de 88 millones de variantes de las cuales 84.7 eran SNPs, 3.6 millones eran indels y 60 mil eran variantes estructurales. Estos recursos incluyen más del 99% de las variantes con una frecuencia superior al 1%. El estudio describe la distribución de la variación genética en la muestra global[54].

A nivel doméstico, otro antecedente relevante en el tema poblacional es el proyecto CHOCOGEN, el cual busca facilitar los estudios genéticos en población afro-colombianas, mediante la caracterización de ancestrías genéticas de la población de Chocó y explorar sus relaciones con determinantes de condición saludable o la enfermedad[84].

Por otra parte, los estudios WES (Whole Exome Sequencing) y de perfiles ómicos (genómica, transcriptómica, proteómica y metabolómicas) también vienen aportando una gran cantidad de información a modo de variantes relacionadas con enfermedades hereditarias o poblacionales, en especial con cáncer y varias enfermedades hereditarias [85].

Se hace necesario explorar nuevas alternativas de enfoques teóricos y predictivos (avenidos de los sistemas dinámicos no lineales) que permitan complementar los estudios anteriores, a fin de identificar en el genoma regiones sensibles o no sensibles a alteraciones. En la literatura, existen pocas referencias, como, por ejemplo, aquellas abordas desde el enfoque multifractal [86] con base en los conteos de repetición de k_{meros} de cromosomas humanos [87] que proponen que las regiones codificantes y no codificantes se diferencian en su multifractalidad.

A pesar que el enfoque de variantes tipo Polimorfismos de Nucleótido Simple, SNP) ha sido de gran utilidad en la identificación de asociaciones con enfermedades y algunos fenotipos, poco a poco se han venido reportando algunas dificultades en determinar estas asociaciones para algunas enfermedades [88], además, aunque el análisis de SNPs ha permitido identificar diferencias críticas que contribuyen a la variación fenotípica de rasgos específicos, solo un grado modesto de variación fenotípica ha sido explicado por polimorfismos de un solo nucleótido. Esto ha llevado a hipótesis más amplias con respecto a la base genética potencial de esta "heredabilidad faltante". De ahí que la variación del número de copias (CNV) se ha propuesto como otro tipo de variación genética que contribuye a la variación fenotípica [89]. Y estas podrían ser detectadas con otros métodos de análisis.

El principal antecedente de crear una teoría predictiva para estudiar el genoma humano, quizás viene de los estudios realizados de la caracterización multifractal del genoma humano[5], el cual reporto los valores multifractales del genoma humano y su relación con elementos repetitivos como las ALU, LINE, islas CpG y otros elementos.

Tomando todos estos hechos en conjunto, se hace necesario explorar nuevas alternativas mediante enfoques disruptivos, teóricos y predictivos poco explorados. El enfoque no lineal ha permitido en poco menos de una década buscar alternativas que bien complementen los estudios anteriores o generen nuevos escenarios de investigación. Todo esto, con el fin de identificar en el genoma asociaciones de regiones sensibles o no sensibles a enfermedades o bien relacionarlos con fenotipos saludables.

El presente trabajo de investigación de las poblaciones humanas colombianas busca desde el punto de vista de los sistemas dinámicos no lineales contribuir a crear una teoría que encaje con muchos hechos existentes y que genere nuevos escenarios por estudiar. A la fecha, existen pocas referencias al respeto, como, por ejemplo, aquellas abordas desde el enfoque multifractal [86] con base en los conteos de repetición de kmeros de cromosomas humanos [87] que proponen que las regiones codificantes y no codificantes del genoma se diferencian en su multifractalidad. No obstante, el principal antecedente en crear una teoría predictiva para estudiar el genoma humano quizás venga de los estudios realizados de la caracterización multifractal del genoma humano [5], el cual reporto por primera vez, una relación directa entre la multifractalidad y los contenidos de algunos elementos repetidos del genoma, como las secuencias ALU e islas CpG. A diferencia de los antecedentes mencionados, el presente estudio no se centra en la identificación de variantes puntuales que caractericen a una población, sino que utiliza el enfoque multifractal para analizar fragmentos o secuencias de ADN (por ventanas) y, desde sus propiedades, extrapolar comportamientos y buscar asociaciones con fenómenos biológicos, como enfermedades o fenotipos saludables en una población.

3.2 MATERIALES Y MÉTODOS

El primer paso al encarar el análisis de secuencias de ADN mediante métodos no lineales es definir su sistema de representación de tal forma que se vea como una señal o una imagen y sea procesable por el algoritmo, para lograr dicha transformación se utilizó la representación del juego del caos [1.3.4].

Con el propósito de efectuar la caracterización multifractal de la población colombiana, se tomó como fuente de datos el proyecto 1000genomes, el cual brinda la información de las variaciones de las muestras frente al genoma referencia GRch37 mediante un conjunto de archivos VCF multimuestra por cada uno de los cromosomas.

El proceso de caracterización inicia generando las secuencias fasta de los 23 pares de cromosomas (excluyendo los Y) de los 2504 individuos a partir del genoma referencia y los archivos vfc correspondientes suministrador por el proyecto 1000genomes. Esto se logra mediante la extracción de las variantes de los archivos multiVCF y aplicándolos individualmente para cada muestra y cada par de cromosomas sobre el genoma referencia. Como resultado de este proceso se obtuvieron 115.184 archivos fasta.

Posteriormente, cada archivo generado se parte en ventanas de 1’000.000 pares base, y se descartan aquellas que contengan un porcentaje mayor al 10% de bases no determinadas (reportadas en los archivos fasta como N). Luego se aplica el proceso de generación de los valores numéricos por medio de la representación del juego del caos (RJC). Dicha representación numérica se utiliza como entrada en la técnica del conteo de cajas para calcular el espectro multifractal hallando el D_q correspondiente a cada q. Los valores de ΔD_q se hallan calculando la diferencia entre D_{-20} y D_{20} y con estos resultados se hace la correspondiente caracterización y análisis. En (la **¡Error! No se encuentra el origen de la referencia.**) se listan las herramientas utilizadas y posteriormente en la (

está el flujo del proceso realizado.

Tabla 5: Lista de herramientas informáticas utilizadas para la caracterización multifractal.

Herramienta	**Utilidad**	
Wget	Herramienta de uso libre para descargar	Se utilizó para realizar la descarga de los archivos de secuencia referencia y

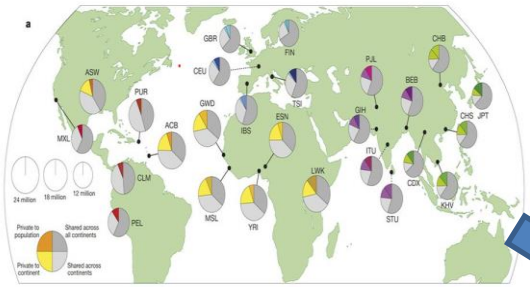
	archivos mediante protocolos HTTP, HTTPS, FTP, FTPS	los archivos VCF del proyecto 1000genomes
Excel	Software de ofimática de Microsoft Corporation	Se utilizó para restaurar las copias de las bases de datos descargadas desde Ensembl que no estaban disponibles online.
Oracle express[14]	Software de base de datos de la compañía Oracle	Se utilizó como repositorio de la base de datos donde se almacenaron las referencias a las diferentes muestras del proyecto 1000genomes así como los cálculos correspondientes de multifractalidad
Pentaho Data Integration CE	Software para realizar procesos de ETL (Extract, transform and load). Se utilizó la versión comunitaria que es de uso gratuito	Se utilizó para descargar y transformar información desde las bases de datos de Ensembl que estaban en mysql y almacenarlas en la base de datos Oracle, también se manejó para automatizar algunos de los pipelines al integrar diversas fuentes de datos, adicionalmente se utilizó para establecer la asociación entre los genes y rutas metabólicas de KEGG
Vcftools	Conjunto de herramientas escritas en PERL para ejecutar tareas comunes con archivos VCF	Se utilizó para procesar los archivos VCF el proyecto 1000genomas y así generar las secuencias fastas que posteriormente serían procesadas para calcular su multifractalidad
SLURM[15]	Slurm es un administrador y agendador de trabajos en ambientes de cluster	Se utilizó para procesar en paralelo la generación de las secuencias fasta y para calcular la multifractalidad de las ventanas de los cromosomas.
Java	Lenguaje de programación	Lenguaje con el que se implementó la generación de RJC, el conteo de cajas y el cálculo multifractal
PowerBI[16]	Herramienta de Microsoft orientada a la generación de reportes	Se utilizó para generar gran parte de las gráficas incluidas en el reporte

[14] Oráculo Aplicaciones Integradas en la nube y servicios de plataforma: www.oracle.com/index.html
[15] Slurm Workload Manager: https://slurm.schedmd.com
[16] Power *Bi*: https://powerbi.microsoft.com/es-es/

RTG TOOLS[17]	Herramienta para comparar y manipular archivos VCF	Se utilizó para extraer la cantidad de SNP de cada muestra

Figura 19: Flujo del proceso de cálculo de los $\Delta D_{q.}$

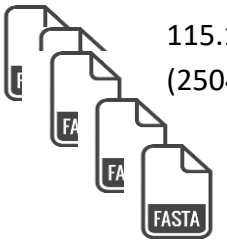
[17] Real Time Genomics Inc: www.realtimegenomics.com/products/rtg-tools



Archivos VCF de 2504 muestras

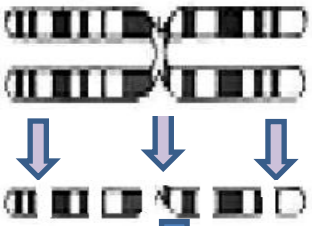
NCBI

Genoma referencia

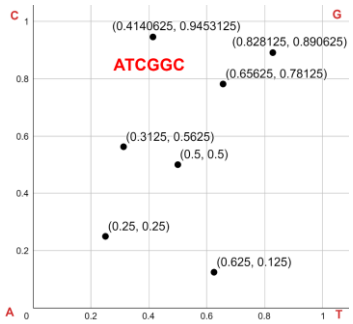


115.184 secuencias
(2504*23*2)

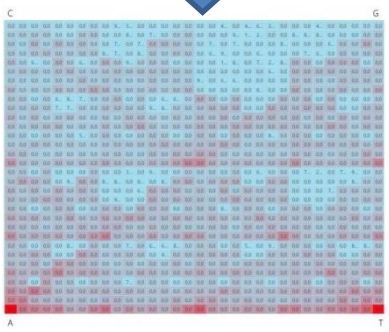
Partición en ventanas de
1’000.000 de PB c/u



Exclusión de ventanas
con más del 10% de N



Cálculo del RJC para
cada ventana



Cálculo del box
counting para cada
ventana y tamaño
de caja

$$D_q(\varepsilon) = \frac{log\left(\sum_i \left(\frac{M_i}{M_0}\right)^q\right)}{\log(\varepsilon)\,(q-1)}$$

Cálculo del espectro multifractal

3.2.1 BASES DE DATOS GENOMICAS

Existen diversas fuentes de información donde se almacenan secuencias genómicas junto con algunas de sus características, utilizando esta información es posible realizar análisis de dichas secuencias y asociarlas con diversas condiciones. Estas son algunas de las bases de datos más utilizadas y que fueron fuente de datos para este trabajo:

1000 Genomas: El **Proyecto 1000 genomas** es una iniciativa para analizar el material genético de mil personas en todo el mundo, con la finalidad de obtener una base de datos que permita estudiar la variabilidad genética humana. El proyecto con un coste de 50 millones de dólares se ha desarrollado en 3 fases, la primera de un año de duración fue en realidad un estudio previo preparatorio, mientras que, en la segunda fase, con una duración de 2 años, se analizaron las secuencias genéticas de un conjunto de 1000 individuos previamente seleccionados, posteriormente se ampliado a 2504 individuos en la tercera fase. En el presente trabajo se ha utilizado la información de estos 2504 individuos ancestrales que comprenden 5 superpoblaciones (AFR, AMR, EAS, EUR, SAS) y 26 poblaciones, incluida la población colombiana (CLM)[18] Figura 20.

Figura 20: Ubicaciones de las poblaciones humanas utilizadas en el proyecto de 1000genomes.

[18] Las muestras colombianas solo son de Medellín

NCBI[19](National Center for Biotechnology Information): Este sitio tiene como objetivo apoyar el desarrollo de la ciencia y la salud suministrando información biomédica y genómica. Tiene la misión de crear sistemas automatizados para almacenar y analizar conocimiento acerca de biología molecular, bioquímica y genética, facilitando el uso de dichas bases de datos y software a la comunidad científica.

Gene Regulation Info[20]: Este sitio está dedicado a la descripción cuantitativa de la regulación génica. Almacena revisiones frecuentemente actualizadas de los recursos en las interacciones proteína-ADN, el posicionamiento de nucleosomas y modificaciones epigenéticas.

ENCODE (Enciclopedia of DNA elements): Es una colaboración internacional de grupos de investigadores fundado por el NHGRI (National Human Genome Research Institute). Su propósito es construir un listado exhaustivo de partes de elementos funcionales del genoma humano, incluyendo elementos que actúan a niveles de RNA y proteínas y elementos regulatorios que controlan las células y las circunstancias en las cuales un gen es activo.

Ensembl[21]: Es un proyecto conformado entre el Instituto Europeo de Bioinformatica y el Wellcome Trust Sanger Institute para ofrecer aparte de bases de datos de genomas, diversos recursos para los investigadores, como documentación y diversas herramientas para el análisis de secuencias.

UCSC[22]: La universidad de California Santa Cruz, junto con otros miembros del consorcio International Human Genome Project participo en el primer ensamblaje del genoma humano, asegurando acceso libre a dicha información mediante la publicación de sus resultados en https://genome.ucsc.edu/ junto con una herramienta grafica de visualización. Desde ese entonces y hasta la fecha, el sitio se ha enriquecido con actualizaciones constantes y herramientas que facilitan el análisis genómico.

[19] NCBI (Centro Nacional de Información Biotecnológica): www.ncbi.nlm.nih.gov
[20] GENE REGULATION LAB: http://generegulation.info/
[21] ENSAMBL GENOMA BROWSER 98: www.ensembl.org/index.html
[22] UNIVERSIDAD DE CALIFORNIA SANTA CRUZ: https://genome.ucsc.edu/

KEGG[23]: Es una base de datos de diversos tipos de recursos que busca ayudar a entender la relación entre sistemas biológicos y sus funciones, como por ejemplo la relación entre los genes y rutas metabólicas.

3.2.2 GENERACIÓN DE SECUENCIAS CONSENSO

El proceso de generación de secuencias consenso consiste en crear archivos fasta a partir de una secuencia referencia y un archivo de especificación de variantes como los archivos VCF. Existen diversos programas y scripts que sirven a este propósito, entre los cuales los más difundidos son las suites BCFTOOLS y SAMTOOLS[24]. Este es un ejemplo de los tipos de scripts utilizados para obtener una secuencia contexto:

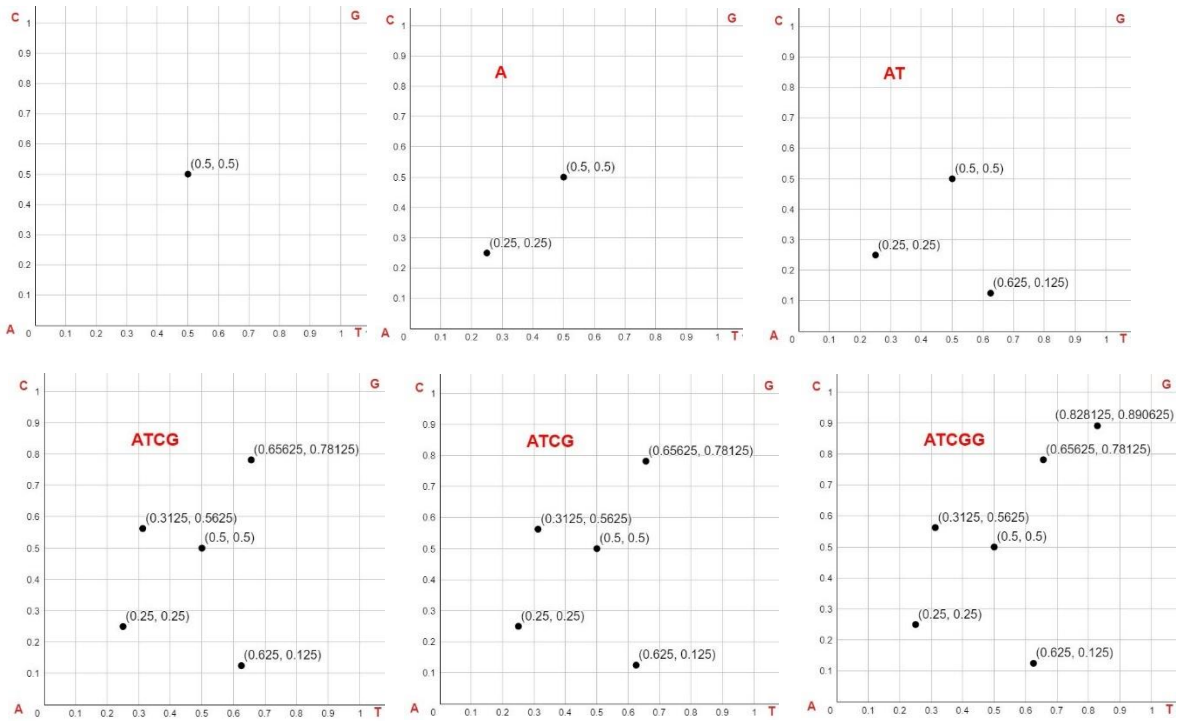
```
#Se extra la región y el individuo de individuo de interés del archivo
# VCF para el cual se desea generar la secuencia contexto:

tabix -h ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.chr17.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz 17:1471000-1472000 | perl vcf-subset -c HG00098 | bgzip -c > HG00098.vcf.gz

#Se indexa el nuevo archivo VCF para que pueda ser utilizado por
#vcf-consensus

tabix -p vcf HG00098.vcf.gz

#Se ejecuta vcf-consensus, el archivo ref.fa sería la secuencia
#referencia en formato fasta

cat ref.fa | vcf-consensus HG00098.vcf.gz > HG00098.fa
```

3.2.3 GENERACIÓN DE RJC

Para cada una de las secuencias de consenso generadas, se ejecutó el proceso de cambio de representación utilizando una implementación propia del RJC. La nueva representación gráfica y/o numérica, se obtiene mediante el mecanismo explicado en 1.3.4. En este ejemplo vamos a utilizar la cadena ATCGG Figura 21.

[23] KEGG ENCICLOPEDIA DE KIOTO GENES Y GENOMAS: www.kegg.jp/
[24] SAMTOOLS: www.htslib.org/

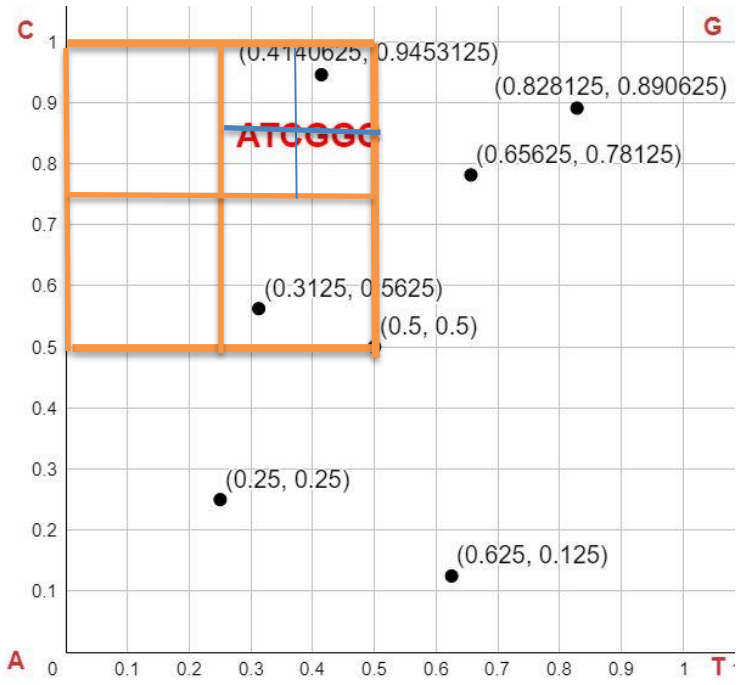
Figura 21: Secuencia de pasos utilizado el RJC para una secuencia de ADN.



La primera letra es la A, se calcula el punto intermedio entre el centro del cuadrado y la esquina en la que se encuentra la A quedando el resultado en (0.25,0.25), la siguiente letra es la T , entonces se calcula el punto medio entre el punto recién generado (0.25,0.25) y la esquina correspondiente a la T, es decir, el punto (1,0), el resultado nos da el punto (0.625,0.125). Se repite el procedimiento de la misma forma hasta finalizar la secuencia.

Este proceso tiene una propiedad matemática interesante: El cuadrante donde se encuentra un punto determinado indica cuales fueron los valores previos de la cadena. Por ejemplo, el punto correspondiente a la C final se encuentra en el cuadrante correspondiente a la C, y si dividimos dicho cuadrante entre cuatro (color naranja), podemos observar que se encontraría en el cuadrante correspondiente a G (la cual es la letra previa en la cadena. Si repetimos el proceso una vez más, nos daremos cuenta de que se encuentra nuevamente en el cuadrante de la G, es decir la letra de dos posiciones atrás Figura 22.

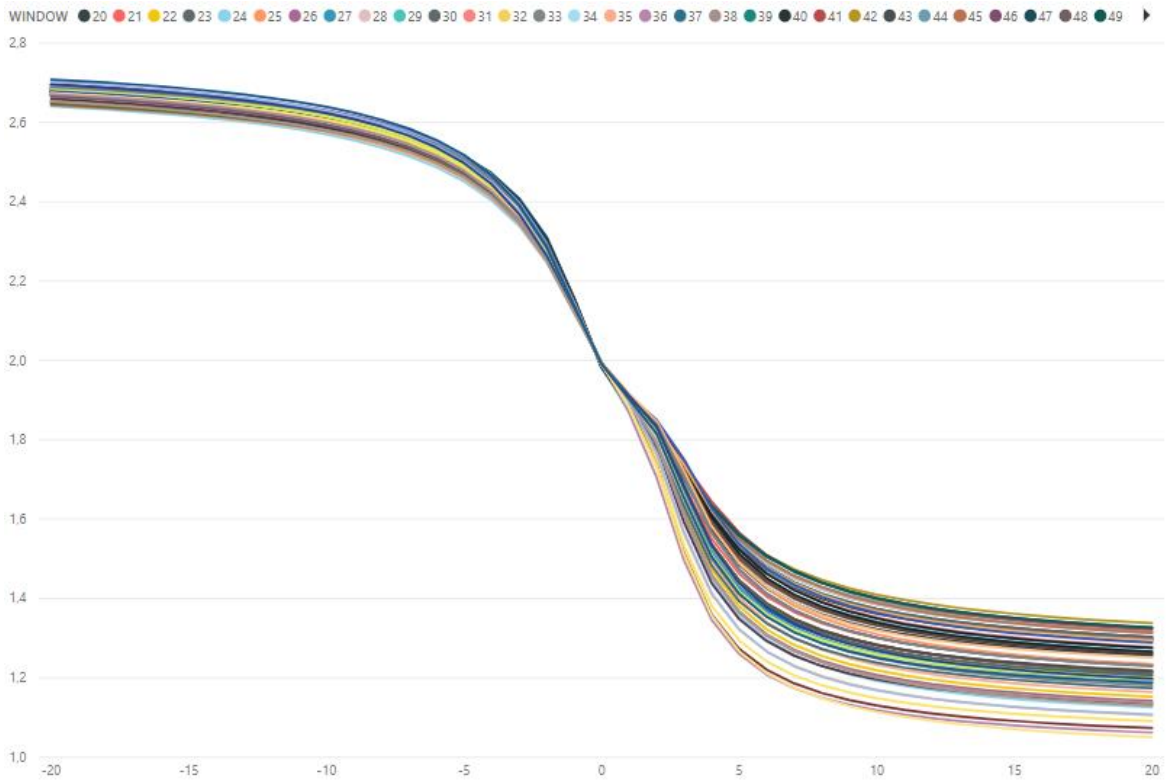
Figura 22: RJC para las primeras 6 bases de la secuencia “ATCGGC…”.



3.2.4 CALCULO DEL ESPECTRO MULTIFRACTAL

Una vez generada la representación RJC de cada una se las secuencias, se procedió a la multifractalidad de cada cromosoma por ventanas de un millón de pares base. En este proceso se utilizó la técnica de BoxCounting junto con el espectro multifractal para calcular los D_q entre los q con valores entre -20 y 20 y finalmente obtener los ΔD_q. Figura 23.

Figura 23: Ejemplo del resultado del cálculo de D_q, para qs entre -20 y 20, para la muestra HG00096, cromosoma 1 ventana 1



¿Cómo se interpreta el ΔD_q? La teoría multifractal establece que: a mayor ΔD_q, mayor contenido de información o mayor multifractalidad e “inestabilidad no lineal” del proceso bajo medición [12]. En el electrocardiograma humano, los espectros multifractales de personas sanas tienen multifractalidades altas, mientras que, en los electrocardiogramas de personas con alguna patología cardiaca, las multifractalidades fueron encontradas bajas o pérdida total de la multifractalidad[90]. La interpretación en este caso pareciera contradictoria, no obstante, lo que sucede es que la inestabilidad no lineal crea “estabilidad en el ritmo cardiaco”. De ahí que el ritmo cardiaco de una persona sana sea totalmente estable (dentro de un rango de inestabilidad caótica). Esto se entiende en el sentido de la teoría de las estructuras disipativas que se alejan del equilibrio, es decir, la aparición de estructuras (o procesos) coherentes, autoorganizados se mantienen en un estado estable lejos del equilibrio [91]. De esta manera, lejos del equilibrio es que el sistema se organiza –estabiliza- produciendo una serie de fenómenos no lineales, a través de las leyes del caos (o aumentos de la MF). Por el contrario, cerca al equilibrio encontramos fenómenos repetitivos o periódicos [90].

En consecuencia, si extendemos estos razonamientos y comportamientos a la secuencia del cromosoma, habrá mayor aperiodicidad de nucleótidos a lo largo de la secuencia del ADN y, por ende, mayor estabilidad en la información genética. De ahí que, en [5](figura 8) se propuso que a mayor MF la secuencia genética tiende a ser más estable, determinista y genética, mientras que a menor MF, la secuencia de ADN es menos estable y tiende a estar más afectada por el medio ambiente y la epigenética. Con base en estas predicciones es que se evalúan e interpretan los cambios de MF extremos de las ventanas de secuencias cromosómicas evaluadas en el presente estudio. Se espera que los valores extremos de altas y bajas multifractalidades encontrados en los genomas a comparar ayuden a identificar aquellas regiones cromosómicas susceptibles de ser estables (genéticamente) e inestables (epigenéticamente) a las fluctuaciones o perturbaciones del medio ambiente. Esto tiene profundas implicaciones a la hora de identificar regiones cromosómicas que contienen genes relacionados con enfermedades humanas, entre otras características y propiedades de los ácidos nucleicos y los genes asociados.

3.3 RESULTADOS CARACTERIZACIÓN MULTIFRACTAL DE LA POBLACIÓN COLOMBIANA

Se calcularon y procesaron 2880 ventanas de 1Mb de longitud repartidas entre los diversos cromosomas de cada muestra, para un total de 14’459.651 valores de ΔD_q, los cuales se utilizaron como base para realizar la caracterización correspondiente.

3.3.1 Promedio total de ΔD_q por cromosoma

Una vez obtenidos los valores de ΔD_q se inició por una caracterización general de los cromosomas, estableciendo el valor promedio para cada uno de ellos (teniendo en cuenta todas las muestras). Este promedio general se obtuvo promediando los valores de cada ventana y posteriormente promediando todas ventanas del mismo cromosoma, Tabla 6.

Tabla 6: Ejemplo con valores del cálculo del valor multifractal(ΔD_q) por cromosoma.

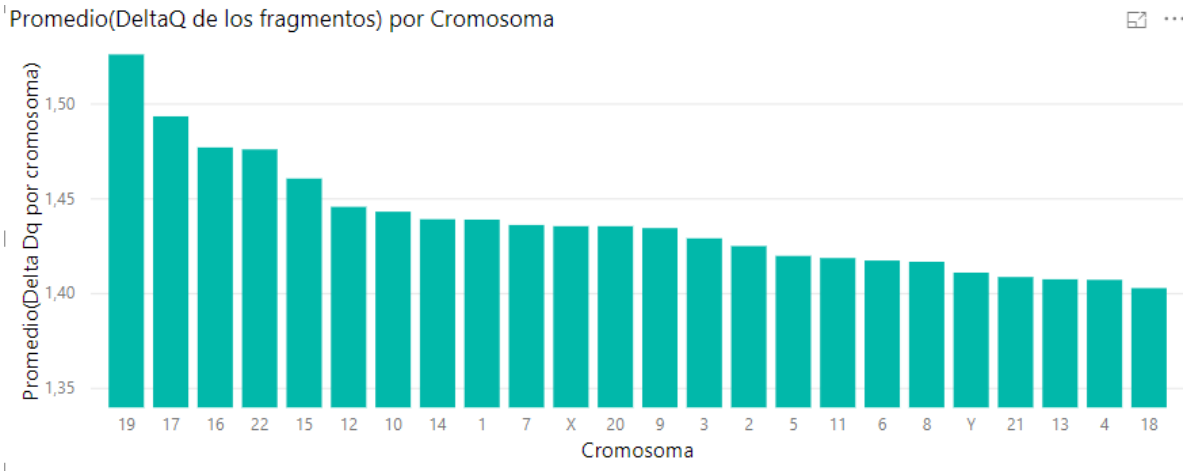
Cromosoma 1	**ventana 1**		**ventana 2**		**ventana 3**		**Ventana N**
	paternal	maternal	paternal	maternal	paternal	maternal	
HG00096	1,31860	1,319240	1,48331	1,49044	1,29326	1,29351	…

HG00097	76 1,31873 13 …	77 1,318893 96 …	3 1,48973 8 …	3 1,48319 73 …	9 1,29314 6 …	56 1,29346 85 …	… ..
Promedio ventana	1,31881263		1,48646754		1,29330733		…

Promedio cromosoma	1,438636

Dichos resultados se asemejan a los obtenidos en [5], donde se procesaron ventanas de 300k pb ,Figura 24. El cromosoma más MF también fue el cromosoma 19, seguido por los cromosomas 17,16 y 22, mientras que, por el contrario, los cromosomas menos MF fueron el 21,13,4, y 8.

Figura 24: Valor promedio de ΔD_q para cada cromosoma teniendo en cuenta todas las muestras.



3.3.2 Comparación de los promedios ΔD_q de todas las muestras vs el ΔD_q del genoma referencia

Posteriormente, se comparó el promedio de ΔD_q de todas las muestras para cada cromosoma contra el ΔD_q de cada cromosoma de secuencia referencia. Se encontró que los valores fueron muy similares entre estos, encontrando pequeñas

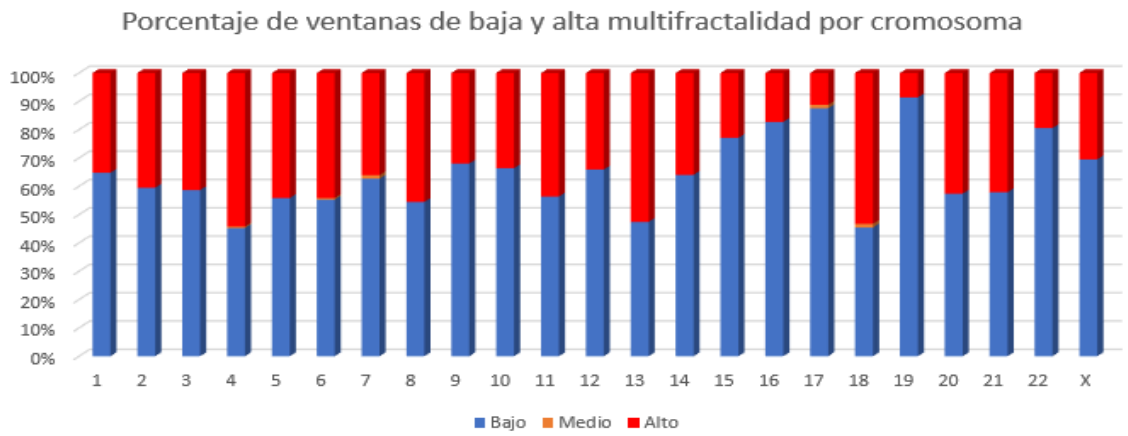
diferencias en los cromosomas 7, 14 y 21 y una diferencia considerable en el cromosoma X, Figura 25.

Figura 25: Promedio de ΔD_q por cromosoma para todas las muestras (en rojo) y para el genoma (en azul).

3.3.3 Porcentaje de ventanas con alta y baja multifractalidad por cromosoma

Luego se calculó el promedio de ventanas por categoría (alta MF $1.4<\Delta D_q$, media MF $1.0<\Delta D_q \leq 1.4$ y baja MF $\Delta D_q \leq 1.0$) que tiene cada cromosoma (como promedio de todas las muestras), Figura 26.

Figura 26: Porcentaje de ventanas de alta media y baja MF por cromosoma.



Se observa que el cromosoma 19 no solo es el más MF en promedio, sino que también tiene la mayor cantidad de ventanas con alta multifractalidad, en contraste, el cromosoma 20 no es el de menor multifractalidad promedio, pero si el que tiene un mayor porcentaje de zonas con baja multifractalidad. Estas observaciones refuerzan lo adecuado del enfoque al trabajar por ventanas ya que se tiene un entendimiento más fino sobre la estructura del cromosoma.

Promedio de ΔD_q por individuo

Posteriormente se generó el promedio de ΔD_q para cada uno de los 2504 individuos (Figura 27), ubicando a los individuos de la misma población de manera contigua y con el mismo color para facilitar el análisis, cada punto en la imagen corresponde a un individuo. En dicho grafico se dio de manera observa la separación de dos grupos: uno en la parte superior y otro en la parte inferior, al explorar las características de cada grupo se detectó que el grupo superior correspondía a hombres y el inferior a mujeres.

Figura 27: Promedio de ΔD_q por individuo

3.3.4 Promedio de ΔD_q por cromosoma por individuo

Para determinar la diferencia entre los grupos de hombres y mujeres se examinó el promedio de ΔD_q por cromosoma para cada individuo (

Figura 28 a Figura 50). Si las líneas tienden a ser horizontales, quiere decir que en ese cromosoma hay poca variación para esa población, por el contrario, si la línea tiende a ser vertical, indica que hay una gran variedad en los valores multifractales de los individuos de la misma población.

Se encontraron diferencias significativas entre los dos grupos (hombre y mujeres) observados en la (Figura 27) y que estas se originaban en el cromosoma X. En los cromosomas 8,13,14,18,19,21 y 22 también se observaron grupos de valores de ΔD_q, sin embargo, estos no correspondían a una división por género. Al no contar con información adicional del individuo aparte de su sexo y población no fue posible encontrar una relación con algún fenómeno biológico.

Figura 28: Promedio de ΔD_q de cada muestra para el cromosoma 1.

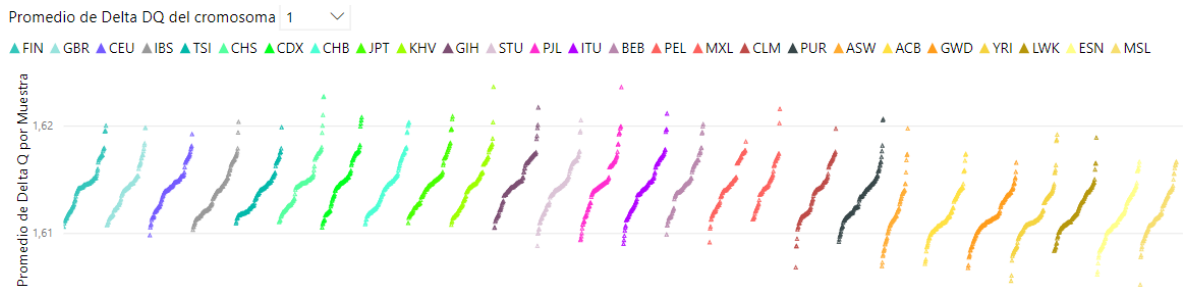


Figura 29: Promedio de ΔD_q de cada muestra para el cromosoma 2.

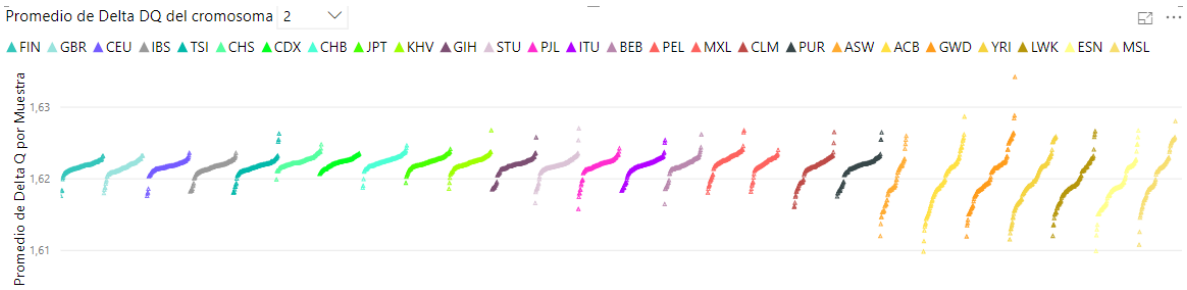


Figura 30: Promedio de ΔD_q de cada muestra para el cromosoma 3.

Figura 31: Promedio de ΔD_q de cada muestra para el cromosoma 4.

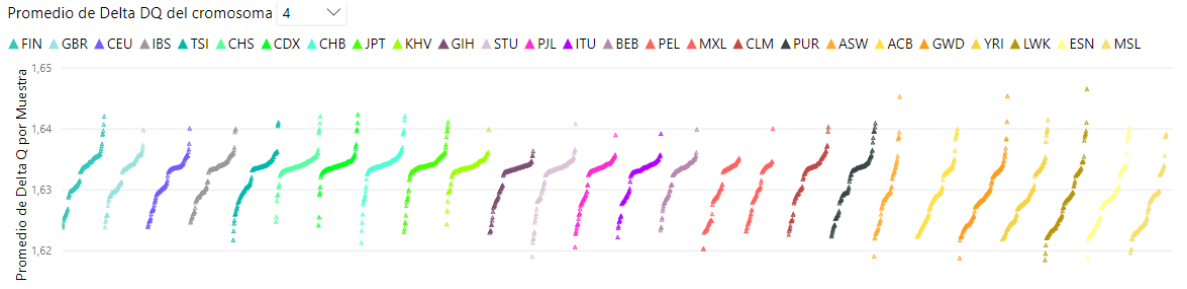


Figura 32: Promedio de ΔD_q de cada muestra para el cromosoma 5.

Figura 33: Promedio de ΔD_q de cada muestra para el cromosoma 6.

Figura 34: Promedio de ΔD_q de cada muestra para el cromosoma 7.

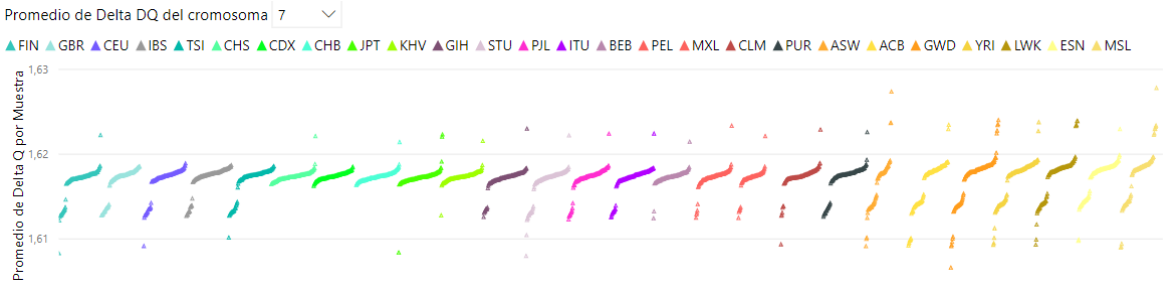


Figura 35: Promedio de ΔD_q de cada muestra para el cromosoma 8.

Figura 36: Promedio de ΔD_q de cada muestra para el cromosoma 9.

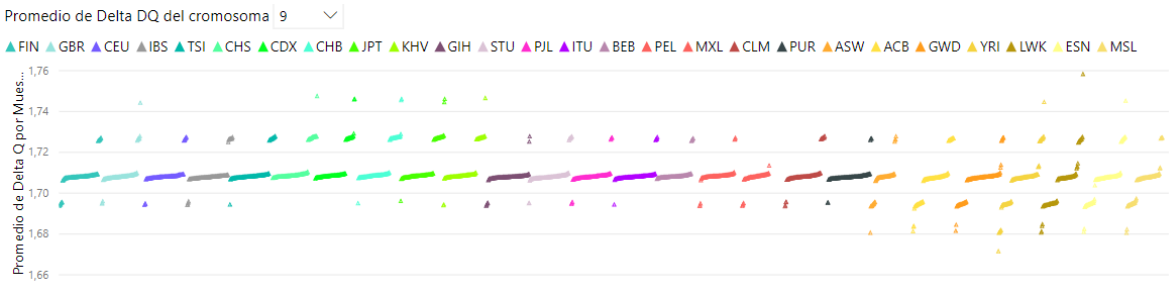


Figura 37: Promedio de ΔD_q de cada muestra para el cromosoma 10.

Figura 38: Promedio de ΔD_q de cada muestra para el cromosoma 11.

Figura 39: Promedio de ΔD_q de cada muestra para el cromosoma 12.

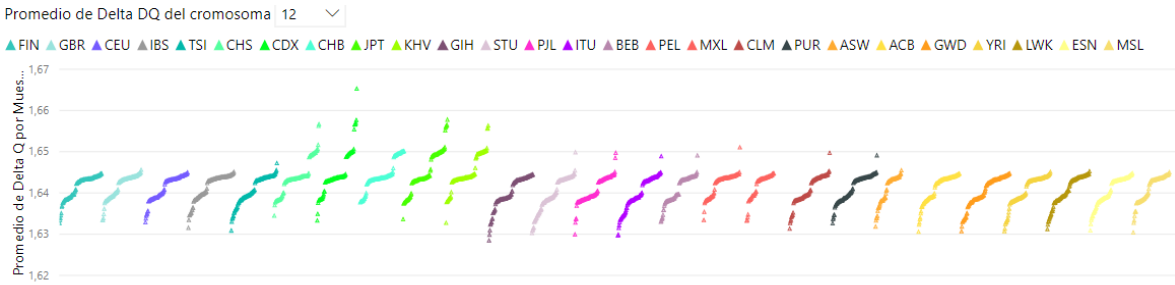


Figura 40: Promedio de ΔD_q de cada muestra para el cromosoma 13.

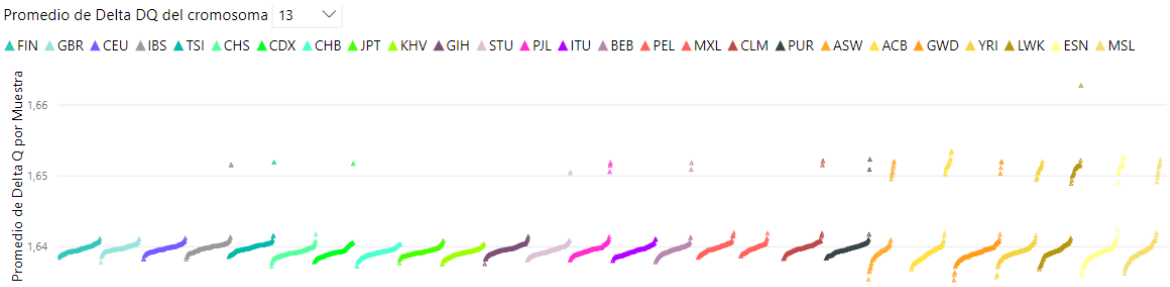


Figura 41: Promedio de ΔD_q de cada muestra para el cromosoma 14.

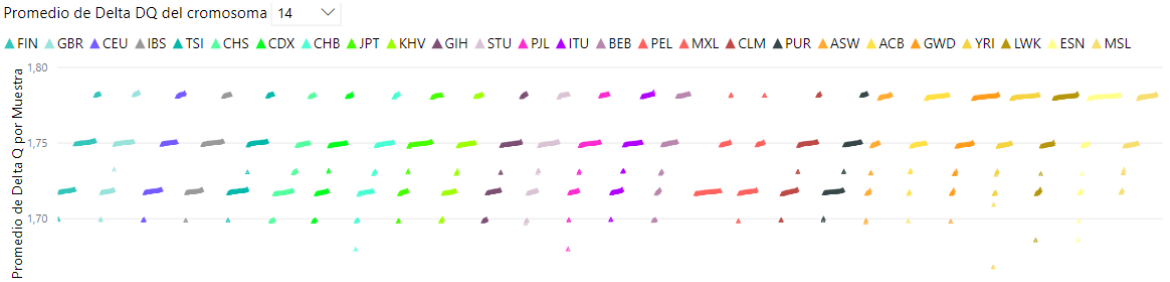


Figura 42: Promedio de ΔD_q de cada muestra para el cromosoma 15.

Figura 43: Promedio de ΔD_q de cada muestra para el cromosoma 16.

Figura 44: Promedio de ΔD_q de cada muestra para el cromosoma 17.

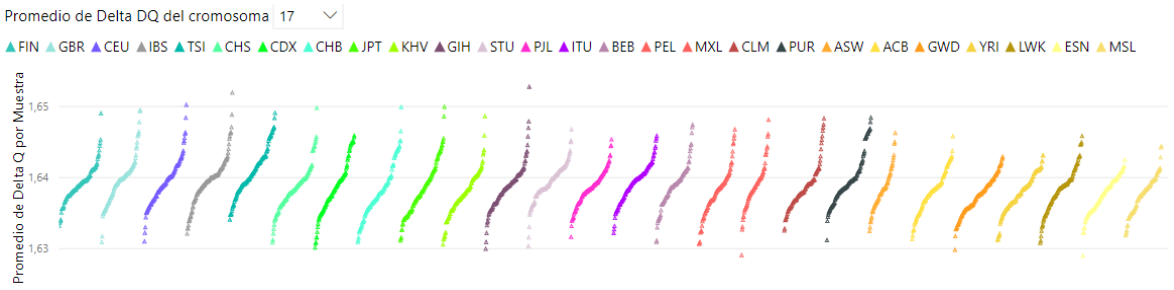


Figura 45: Promedio de ΔD_q de cada muestra para el cromosoma 18.

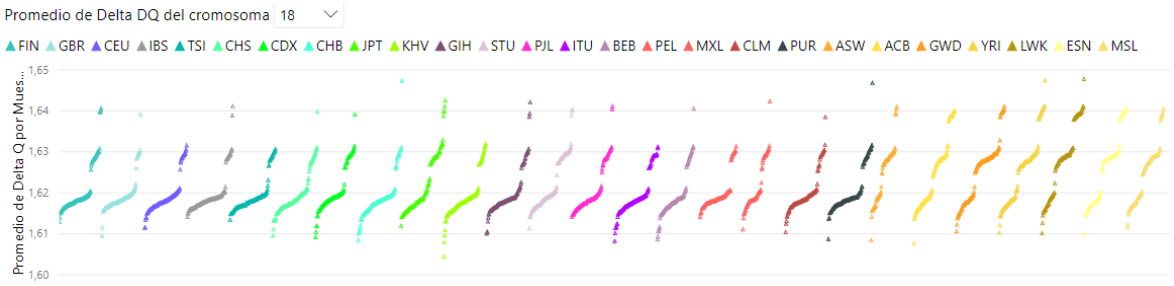


Figura 46: Promedio de ΔD_q de cada muestra para el cromosoma 19.

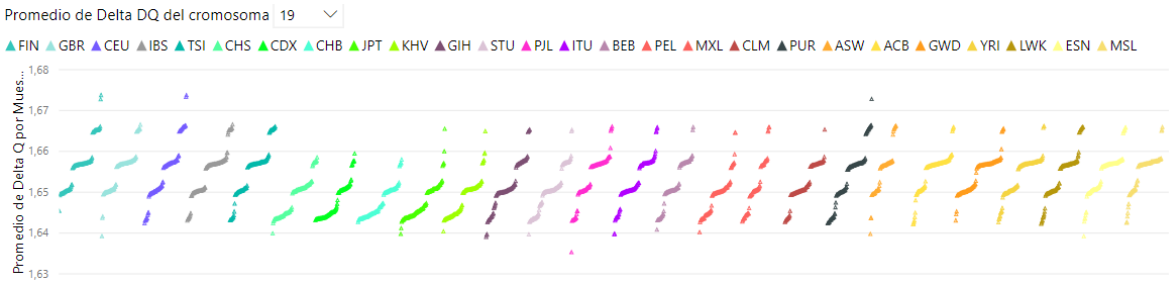
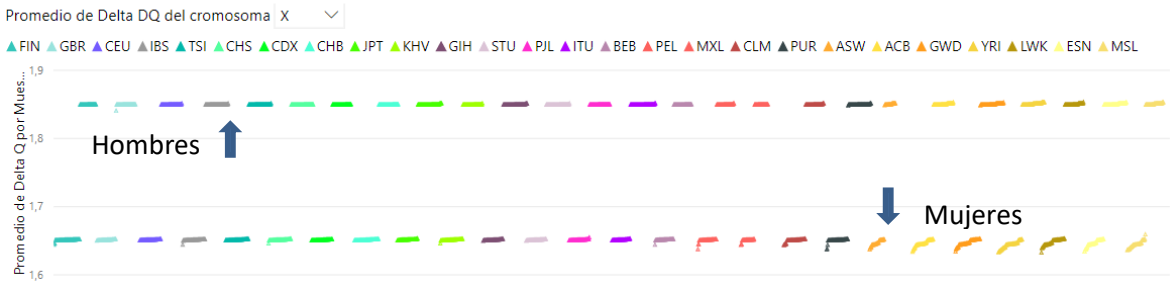


Figura 47: Promedio de ΔD_q de cada muestra para el cromosoma 20.

Figura 48: Promedio de ΔD_q de cada muestra para el cromosoma 21.

Figura 49: Promedio de ΔD_q de cada muestra para el cromosoma 22.

Figura 50: Promedio de ΔD_q de cada muestra para el cromosoma X.

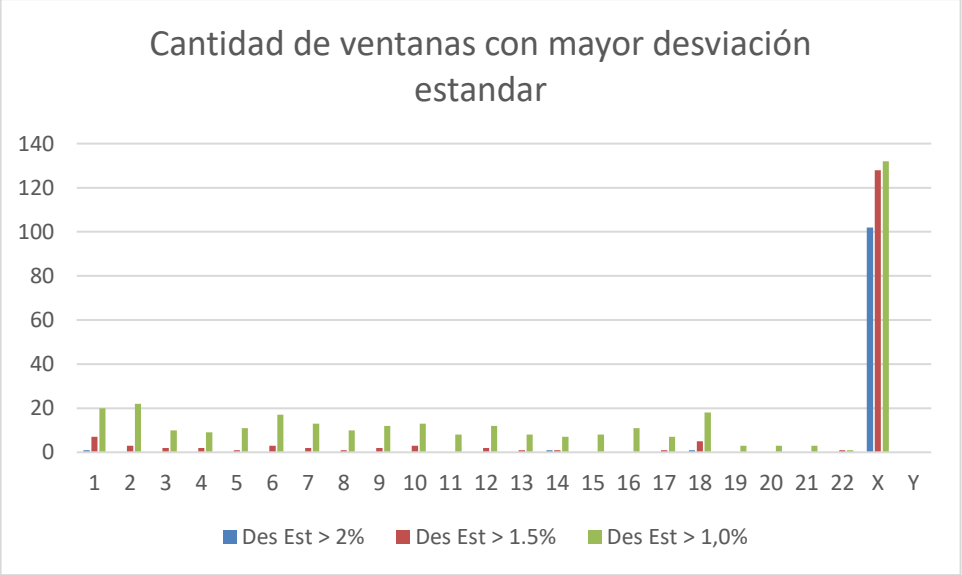


En la Figura 50 se observa la evidente separación de los muestras de hombres en la partes superior y mujeres en la parte inferior en cuanto al cromosoma X, el aporte de este cromosoma es el que hace que en el gráfico de multifractalidad general por individuo (Figura 27) se observen dichas diferencias.

3.3.5 Análisis por ventanas cromosómicas con mayor variación

Para determinar el grado de variación de los cromosomas se hizo un análisis por ventanas de 1Mb de longitud. Se identificaron cuales ventanas eran más variables mediante el cálculo de la desviación estándar teniendo en cuenta los valores de las distintas muestras. Se buscó identificar aquellas zonas con una mayor diversidad, encontrando 105 ventanas con una desviación estándar mayor del 2%. La gran mayoría de estas ventanas se encontraron en el cromosoma X. Las ventanas de otros cromosomas solo se observan cuando se tienen en cuenta aquellas que tienen la desviación estándar mayor que uno, Figura 51. Se puede ver el listado de las ventanas en el material suplementario.

Figura 51: Ventanas con multifractalidad con mayor desviación estándar.



3.3.6 Caracterización de la población colombiana

Con el objeto de caracterizar la población colombiana se determinaron los valores máximos, mínimo y promedio de ΔD_q por cada una de las 26 poblaciones, encontrando los siguientes valores, (

.

Tabla 7: Valores mínimos, máximos y promedio de ΔD_q de ventanas de Mb de longitud en la población colombiana

Cromosoma	Min	Max	Promedio
1	1,08894129	1,66522823	1,43835538
2	1,0726616	1,64990845	1,42433354
3	1,23883721	1,61769258	1,4280962
4	0,88155125	1,770757	1,40624894
5	1,16556739	1,61641027	1,41885831
6	0,92878601	1,63167955	1,41626638
7	0,97144038	1,67855212	1,43509443
8	1,12960858	1,58401581	1,41540745
9	1,15343501	1,63066313	1,43246234
10	1,24624569	1,65101384	1,44326266
11	1,22759503	1,64031513	1,41745905
12	1,10699245	1,65347204	1,44423929
13	1,1439454	1,59348235	1,40524069
14	1,03338408	1,62721927	1,43588991
15	1,26545913	1,63875962	1,45852796
16	1,12907411	1,62565889	1,47531647
17	0,98484568	1,67200542	1,49174457
18	0,68198808	1,57726236	1,40016202
19	1,19032416	1,64677534	1,52391447
20	1,15165719	1,66003779	1,43201435
21	1,045814	1,67337883	1,40295085
22	1,2447476	1,6348626	1,47113767
X	1,14352667	1,63077841	1,4319721

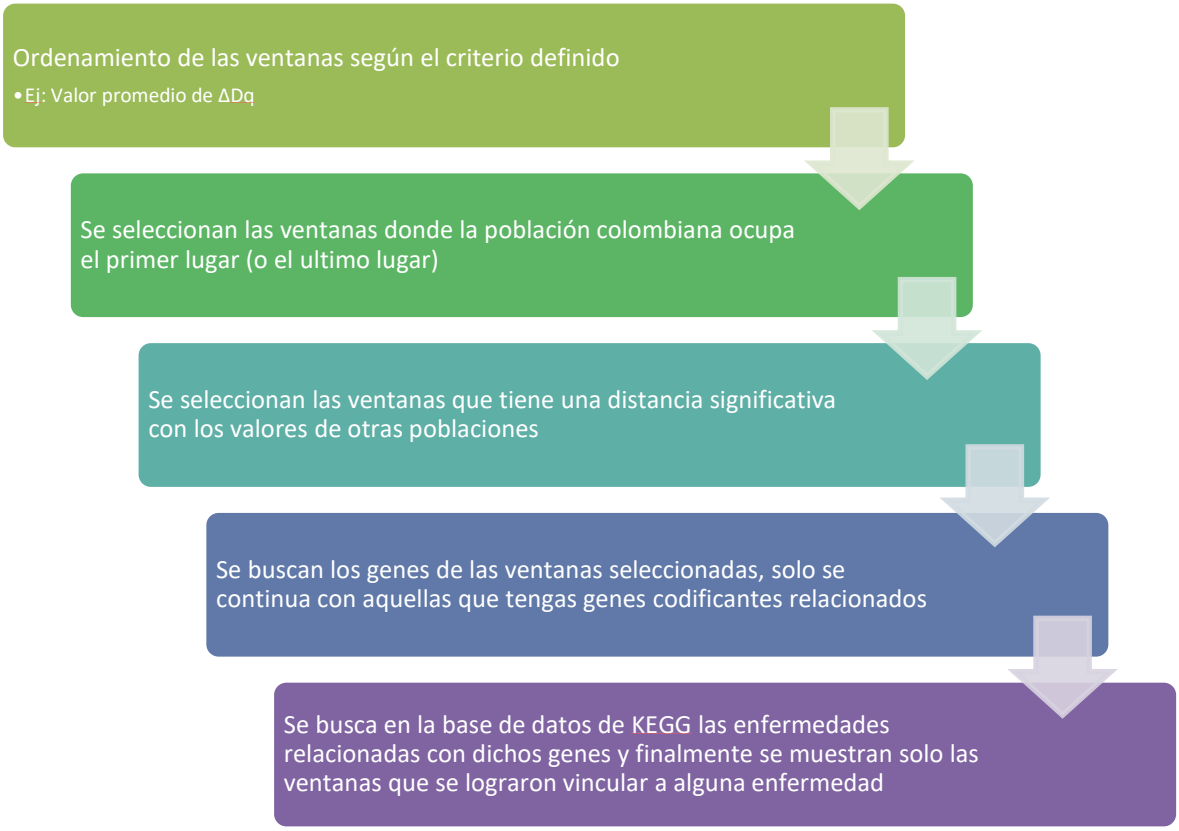
Se encontró que el valor mínimo para una ventana fue de 0,6819 en el cromosoma 18 y el valor máximo de 1,6785 en el cromosoma 7.

3.3.7 Colombia vs. Otras poblaciones humanas

A partir de los datos obtenidos se inició un estudio comparativo entre los colombianos y las otras poblaciones tomando los valores de la multifractalidad de las diferentes ventanas de los cromosomas. Primero se efectuó una caracterización de las ventanas de las muestras colombianas identificando su **amplitud de variación** (valor máximo menos valor mínimo), posteriormente se identificaron las ventanas con mayor y menor desviación estándar. Finalmente se hizo una comparación con las ventanas de las otras poblaciones incluidas en el

estudio, determinando aquellas donde la fractalidad promedio de las muestras colombianas es superior o inferior a todas las demás poblaciones Figura 52.

Figura 52: Pasos para la selección de ventanas con valores extremos para la población colombiana.



Rangos de amplitud por ventana

En este enfoque se obtiene el rango de variación por ventana mediante la diferencia entre el valor máximo y mínimo de las muestras de cada grupo en cada ventana en cada cromosoma. Se encontró que, los colombianos son los de menor variación en 12 de ellas. Estas ventanas solo sugieren aquellas características donde No hay variedad entre los individuos de la población, es decir que son regiones conservadas, pero no necesariamente características diferenciadoras de la población.

Tabla 8: Ventanas de muestra colombianas de menor rango de variación frente a los valores de otras poblaciones

C	W	CLM	EUR	SAS	EAS	AFR	AMR
11	65	0,00539	0,00681	0,01064	0,01309	0,01821	0,00736
13	23	0,0099	0,0341	0,03826	0,039	0,03751	0,03888
13	40	0,01086	0,01407	0,01455	0,01298	0,01784	0,01452
13	41	0,00799	0,00951	0,01169	0,01147	0,01214	0,01107
13	95	0,00647	0,00981	0,00946	0,00981	0,01146	0,00867
15	81	0,00639	0,00844	0,00897	0,00915	0,03493	0,00793
16	25	0,03434	0,03706	0,03711	0,03761	0,03841	0,03744
16	29	0,0126	0,01375	0,01809	0,03082	0,01905	0,01509
18	3	0,03474	0,03722	0,03723	0,03714	0,0391	0,03795
18	14	0,03559	0,04002	0,05165	0,05329	0,05555	0,05392
18	30	0,00483	0,00736	0,03461	0,03607	0,00933	0,007
18	33	0,00426	0,00621	0,00607	0,00607	0,00844	0,00673

Se encontraron que los cromosomas X, 2 y 13 tienen 4 ventanas, los cromosomas 7, 11 y 16, 5 ventanas y el cromosoma 18, 8 ventanas, (

). La Tabla 9 muestra las enfermedades asociadas a cada ventana.

Tabla 9: Enfermedades relacionadas con las ventanas donde los colombianos tienen menor variación respecto a otras poblaciones.

C	W	GEN	ENFERMEDAD
X	16	PIGA	Inherited glycosylphosphatidylinositol deficiencies,Multiple congenital anomalies-hypotonia-seizures syndrome,Paroxysmal nocturnal hemoglobinuria
X	16	ZRSR2	Myelodysplastic síndrome
X	129	OCRL	Dent disease,Lowe syndrome; Oculocerebrorenal Dystrophy (OCRL)
X	129	ZDHHC9	Syndromic X-linked mental retardation
1	225	WDR26	Skraban-Deardorff síndrome
11	65	CAPN1	Hereditary spastic paraplegia

11	65	MEN1	Adrenal carcinoma,Bilateral macronodular adrenal hyperplasia,Carcinoid,Cushing syndrome,Medulloblastoma,Multiple endocrine neoplasia syndrome; Wermer syndrome; Sipple syndrome,Pancreatic neuroendocrine tumor,Pituitary adenomas,Primary hyperparathyroidism; Familial hyperparathyroidism,Zollinger-Ellison syndrome
11	65	PYGM	Glycogen storage disease,Glycogen storage disease type V; McArdle disease,Muscle glycogen storage disease
11	65	RASGRP2	Bleeding disorder platelet-type
11	65	SLC22A12	Renal hypouricemia
13	23	FGF9	Multiple synostosis síndrome
13	40	FREM2	Fraser síndrome
13	41	COG6	Congenital disorders of glycosylation type II
13	95	GPC6	Heparan sulfate proteoglycan gene defects,Omodysplasia
15	81	ARNT2	Webb-Dattani syndrome; Hypothalamic insufficiency-secondary microcephaly-visual impairment-urinary anomalies syndrome
15	81	FAH	Tyrosinemia
16	25	TNRC6A	Familial adult myoclonic epilepsy; Benign adult familial myoclonic epilepsy
16	29	ATP2A1	Brody myopathy
16	29	CD19	Common variable immunodeficiency
16	29	CLN3	Batten disease; Spielmeyer-Vogt disease; Juvenile neuronal ceroid lipofuscinoses,Neuronal ceroid lipofuscinosis
16	29	TUFM	Combined oxidative phosphorylation deficiency
18	3	LPIN2	Chronic recurrent multifocal osteomyelitis; Majeed síndrome
18	3	SMCHD1	Facioscapulohumeral muscular dystrophy
18	14	MC2R	Adrenal carcinoma,Bilateral macronodular adrenal hyperplasia,Cushing syndrome,Familial glucocorticoid deficiency
18	30	DSG2	Arrhythmogenic right ventricular cardiomyopathy
18	30	RNF125	Tenorio síndrome
18	30	TTR	Familial amyloidosis,Familial carpal tunnel

			síndrome
18	33	DTNA	Left ventricular noncompaction
18	33	MAPRE2	Congenital symmetric circumferential skin creases; Kunze-Riehm syndrome; Michelin tire baby syndrome
2	12	LPIN1	Acute recurrent myoglobinuria
2	240	PER2	Familial advanced sleep phase síndrome
2	240	TRAF3IP1	Senior-Loken syndrome
2	240	TWIST2	Ablepharon-macrostomia syndrome,Barber-Say syndrome,Focal facial dermal dysplasia
3	28	EOMES	Polymicrogyria
5	159	IL12B	IFN-gamma/IL-12 axis; Medelian susceptibility to mycobacterial disease (MSMD)
6	109	OSTM1	Osteopetrosis
6	109	SEC63	Polycystic liver disease
7	7	RAC1	Autosomal dominant mental retardation
7	81	CD36	CD36 deficiency; Platelet glycoprotein IV deficiency,Coronary artery disease
7	95	COL1A2	Ehlers-Danlos syndrome arthrochalasia type,Ehlers-Danlos syndrome cardiac valvular type,Osteogenesis imperfecta,Osteoporosis
7	95	PON1	Diabetic retinopathy
7	95	SGCE	Primary dystonia
8	87	CA2	Combined proximal and distal renal tubular acidosis,Osteopetrosis,Renal tubular acidosis

A continuación, se determinaron las ventanas con mayor variación. Las ventanas donde Colombia tiene el mayor rango de variación (máximo valor – mínimo valor) en comparación al resto de poblaciones son 16. Se encontraron para los cromosomas 2 y 16, 8 ventanas, y para el cromosoma 12, 22 ventanas, (Tabla 10). En la Tabla 11 se relacionan las enfermedades vinculadas a dichas ventanas.

Tabla 10: Ventanas de muestra colombianas de mayor variación frente a los valores de otras poblaciones

C	W	CLM	EUR	SAS	EAS	AFR	AMR
11	129	0,0577	0,04401	0,04434	0,04186	0,05606	0,04219
12	9	0,06382	0,03444	0,03469	0,06448	0,03731	0,03498
12	16	0,01187	0,0103	0,01039	0,01086	0,01258	0,01052

12	53	0,03569	0,02754	0,0277	0,02539	0,02666	0,03475
12	99	0,01025	0,00804	0,00847	0,00815	0,00894	0,00764
13	29	0,05124	0,03679	0,03535	0,04972	0,02241	0,00665
14	96	0,06646	0,06637	0,06623	0,03561	0,06668	0,03432
15	35	0,0267	0,02636	0,02603	0,02682	0,02634	0,02454
15	80	0,06835	0,02965	0,04357	0,02504	0,06731	0,02522
16	16	0,01587	0,01584	0,0114	0,01512	0,01601	0,01515
16	57	0,03484	0,03339	0,03374	0,03471	0,03439	0,03302
16	67	0,03622	0,03599	0,03622	0,03519	0,03684	0,03616
17	44	0,03122	0,02137	0,00644	0,00521	0,03105	0,02024
18	21	0,04952	0,04967	0,04855	0,04978	0,05064	0,0491
18	51	0,02591	0,01743	0,01671	0,01644	0,02381	0,01879
20	3	0,02555	0,02505	0,02546	0,02293	0,01315	0,02485

Tabla 11: Enfermedades relacionadas con las ventanas donde los colombianos tienen mayor variación respecto a otras poblaciones

C	W	GEN	ENFERMEDAD
X	47	RP2	Retinitis pigmentosa
X	132	FRMD7	Congenital motor nystagmus (CMN); Idiopathic congenital nystagmus (ICN)
X	137	GPR101	Acromegaly,Cushing syndrome,Excessive secretion of growth hormone,Pituitary adenomas,Pituitary gigantism
X	137	ZIC3	Congenital heart defects, multiple type,Heterotaxy; Situs ambiguus,VACTERL/VATER association
X	148	AFF2	Syndromic X-linked mental retardation
X	148	FMR1	Fragile X syndrome,Fragile X tremor/ataxia syndrome,Premature ovarian failure
1	154	GATAD2B	Autosomal dominant mental retardation
1	154	LOR	Vohwinkel síndrome
11	129	FLI1	Bleeding disorder platelet-type,Ewing sarcoma
11	129	KCNJ1	Bartter síndrome
11	129	KCNJ5	Long QT syndrome,Primary aldosteronism
12	9	AICDA	Hyper IgM syndromes, autosomal recessive type
12	9	MFAP5	Familial thoracic aortic aneurysm and dissection; Aortic aneurysm familial thoracic type (AAT)

12	9	NECAP1	Early infantile epileptic encephalopathy; Ohtahara síndrome
12	16	EPS8	Deafness, autosomal recessive
12	16	MGP	Keutel síndrome
12	16	PDE6H	Achromatopsia; Rod monochromacy,Cone-rod dystrophy and cone dystrophy
12	16	PTPRO	Nephrotic síndrome
12	53	ACVRL1	Hereditary hemorrhagic telangiectasia; Osler disease,Pulmonary arterial hypertension
12	53	KRT5	Dowling-Degos disease,Epidermolysis bullosa simplex
12	53	KRT6A	Pachyonychia congénita
12	53	KRT6B	Pachyonychia congénita
12	53	KRT6C	Non-epidermolytic palmoplantar keratoderma
12	53	KRT71	Hereditary hypotrichosis simplex
12	53	KRT74	Ectodermal dysplasia, hair-nail type,Hereditary hypotrichosis simplex,Woolly hair
12	53	KRT75	Pseudofolliculitis barbae
12	53	KRT81	Monilethrix
12	53	KRT83	Erythrokeratodermia variabilis,Monilethrix
12	53	KRT85	Ectodermal dysplasia, hair-nail type
12	53	KRT86	Monilethrix
12	53	SCN8A	Early infantile epileptic encephalopathy; Ohtahara síndrome
12	99	SLC25A3	Mitochondrial phosphate carrier deficiency
12	99	TMPO	Dilated cardiomyopathy
13	29	CDX2	Gastric cancer
13	29	FLT3	Acute myeloid leukemia
13	29	PDX1	Maturity onset diabetes of the young (MODY),Pancreatic agenesis,Permanent neonatal diabetes mellitus
13	29	POLR1D	Treacher Collins síndrome
14	96	DICER1	Cushing syndrome,Pituitary adenomas
15	35	NOP10	Dyskeratosis congénita
15	35	SLC12A6	Agenesis of the corpus callosum with peripheral neuropathy (ACCPN)
15	80	CTSH	Type 1 diabetes mellitus
16	16	MYH11	Familial thoracic aortic aneurysm and dissection; Aortic aneurysm familial thoracic type (AAT),Patent ductus arteriosus
16	16	NDE1	Lissencephaly,Microhydranencephaly

16	57	BBS2	Bardet-Biedl syndrome,Retinitis pigmentosa
16	57	GNAO1	Early infantile epileptic encephalopathy; Ohtahara síndrome
16	57	NUP93	Nephrotic síndrome
16	57	SLC12A3	Gitelman síndrome
16	67	BEAN1	Spinocerebellar ataxia (SCA)
16	67	TK2	Autosomal recessive progressive external ophthalmoplegia,Mitochondrial DNA depletion syndrome
17	44	PLEKHM1	Osteopetrosis
18	21	RBBP8	Seckel síndrome
18	51	DCC	Cancer of the anal canal,Colorectal cancer,Congenital mirror movements
2	39	CYP1B1	Anterior segment dysgenesis,Primary congenital glaucoma; Glaucoma 3
2	45	ABCG5	Macrothrombocytopenia,Sitosterolemia
2	45	ABCG8	Gallbladder disease,Macrothrombocytopenia,Sitosterolemia
2	45	DYNC2LI1	Short-rib thoracic dysplasia
2	45	LRPPRC	Cytochrome c oxidase (COX) deficiency; Mitochondrial complex IV deficiency (MT-C4D),Leigh syndrome
2	45	PREPL	Congenital myasthenic síndrome
2	45	SLC3A1	Cystinuria
2	63	FAM161A	Retinitis pigmentosa
20	3	IDH3B	Retinitis pigmentosa
20	3	NOP56	Spinocerebellar ataxia (SCA)
20	3	SNRPB	Cerebrocostomandibular síndrome
20	3	TGM3	Uncombable hair syndrome; Spun glass hair syndrome; Pili trianguli et canaliculi
20	3	TGM6	Spinocerebellar ataxia (SCA)
3	5	ITPR1	Spinocerebellar ataxia (SCA)
3	5	SUMF1	Multiple sulfatase deficiency,Sphingolipidosis
3	54	CACNA1D	Primary aldosteronism
3	54	PRKCD	Autoimmune lymphoproliferative syndromes
3	54	RFT1	Congenital disorders of glycosylation type I
3	124	MYLK	Familial thoracic aortic aneurysm and dissection; Aortic aneurysm familial thoracic type (AAT)
4	53	SGCB	Limb-girdle muscular dystrophy,Sarcoglycanopathies
5	154	HAND1	Hypoplastic left heart síndrome

6	110	ARMC2	Spermatogenic failure
6	110	CD164	Deafness, autosomal dominant
6	110	ZBTB24	Immunodeficiency-centromeric instability-facial anomalies síndrome
6	118	RFX6	Mitchell-Riley síndrome
6	118	ROS1	Non-small cell lung cancer
6	140	CITED2	Atrial septal defect,Ventricular septal defect
6	140	REPS1	Neurodegeneration with brain iron accumulation
7	27	HNRNPA2B1	Inclusion body myopathy with Paget disease of bone and frontotemporal dementia
7	27	SNX10	Osteopetrosis
7	94	CALCR	Osteoporosis
7	141	BRAF	Cardiofaciocutaneous syndrome,Langerhans cell histiocytosis,Leopard syndrome,Melanoma,Noonan syndrome,Noonan syndrome and related disorders,Thyroid cancer
8	34	TTI2	Autosomal recessive mental retardation
8	76	GDAP1	Charcot-Marie-Tooth disease; Hereditary motor and sensory neuropathy; Peroneal muscular atrophy
8	76	JPH1	Charcot-Marie-Tooth disease; Hereditary motor and sensory neuropathy; Peroneal muscular atrophy
9	121	TLR4	Age-related macular degeneration
9	133	TOR1A	Primary dystonia

Desviación estándar por ventanas

Se calcularon las desviaciones estándar de cada ventana teniendo en cuenta los valores de ΔDq. Estas son las ventanas donde las muestras colombianas tenían una menor desviación estándar, (Tabla 12) y enfermedades asociadas, (Tabla 13).

Tabla 12: Ventanas de muestras colombianas con menor desviación estándar frente a otras poblaciones.

C	W	CLM	EUR	SAS	EAS	AFR	AMR
X	24	0,01465	0,015	0,01526	0,015	0,01486	0,01488
1	146	0,00072	0,0008	0,00088	0,0012	0,001	0,00082
11	65	0,00113	0,0012	0,00158	0,00144	0,00137	0,00125
12	112	0,00117	0,0024	0,00135	0,00424	0,00143	0,00185

13	95	0,00124	0,0014	0,00146	0,00143	0,00171	0,00155
15	50	0,00097	0,0012	0,0011	0,00105	0,00125	0,00109
19	1	0,00134	0,0016	0,0018	0,00148	0,00154	0,00154
3	15	0,00414	0,0053	0,00551	0,00527	0,00481	0,00624
3	122	0,00138	0,0015	0,00162	0,00187	0,00167	0,00148
3	173	0,00122	0,0015	0,00156	0,00149	0,00165	0,00134
4	56	0,00193	0,0021	0,00226	0,00215	0,00237	0,00212
4	104	0,0018	0,0021	0,0022	0,00206	0,00226	0,00217
4	111	0,00134	0,0016	0,00169	0,00152	0,00157	0,00149
7	7	0,00067	0,0009	0,00126	0,00116	0,00143	0,00083
7	158	0,00185	0,0022	0,00203	0,00219	0,0026	0,00217

Tabla 13: Enfermedades relacionadas con las ventanas donde los colombianos tienen menor variación estándar respecto a otras poblaciones

C	W	GEN	ENFERMEDAD
X	24	PTCHD1	Autism; Autistic spectrum disorder; Pervasive developmental disorder
X	24	SAT1	Keratosis follicularis spinulosa decalvans
1	146	HFE2	Hemochromatosis
11	65	CAPN1	Hereditary spastic paraplegia
11	65	MEN1	Adrenal carcinoma,Bilateral macronodular adrenal hyperplasia,Carcinoid,Cushing syndrome,Medulloblastoma,Multiple endocrine neoplasia syndrome; Wermer syndrome; Sipple syndrome,Pancreatic neuroendocrine tumor,Pituitary adenomas,Primary hyperparathyroidism; Familial hyperparathyroidism,Zollinger-Ellison síndrome
11	65	PYGM	Glycogen storage disease,Glycogen storage disease type V; McArdle disease,Muscle glycogen storage disease
11	65	RASGRP2	Bleeding disorder platelet-type
11	65	SLC22A12	Renal hypouricemia
12	112	ATXN2	Spinocerebellar ataxia (SCA)
12	112	MYL2	Hypertrophic cardiomyopathy
12	112	SH2B3	Type 1 diabetes mellitus
12	112	TCTN1	Joubert síndrome
13	95	GPC6	Heparan sulfate proteoglycan gene defects,Omodysplasia
15	50	CEP152	Primary microcephaly,Seckel síndrome
19	1	CFD	Alternative complement pathway component

			defects
19	1	ELANE	Neutropenic disorders
19	1	KISS1R	Central precocious puberty,Hypogonadotropic hypogonadism,Precocious puberty
3	15	CCDC174	Infantile hypotonia with psycHomotor retardation and characteristic facies
3	15	TMEM43	Arrhythmogenic right ventricular cardiomyopathy
3	15	XPC	Disorders of nucleotide excision repair,Xeroderma pigmentosum
3	122	CASR	Bartter syndrome,Calcium sensing receptor (CASR) related disease,Familial hypocalciuric hypercalcemia,Idiopathic generalized epilepsies,Neonatal hyperparathyroidism,Parathyroid carcinoma,Primary hyperparathyroidism; Familial hyperparathyroidism
3	122	ILDR1	Deafness, autosomal recessive
3	122	IQCB1	Nephronophthisis,Senior-Loken síndrome
3	173	SPATA16	Globozoospermia; Round-headed spermatozoa,Spermatogenic failure
4	56	KDR	Infantile hemangioma
4	56	KIT	Acute myeloid leukemia,Anaphylaxis,Breast cancer,Gastrotintestinal stromal tumor,Mast-cell leukemia,Piebaldism
4	56	PDGFRA	Chronic eosinophilic leukemia,Gastrotintestinal stromal tumor,Glioma,Hypereosinophilic síndrome
4	104	CISD2	Wolfram syndrome; DIDMOAD síndrome
4	104	MANBA	Glycoproteinoses,beta-Mannosidosis
4	104	NFKB1	Common variable immunodeficiency
4	104	SLC39A8	Congenital disorders of glycosylation type II
4	111	CFI	Age-related macular degeneration,Atypical hemolytic uremic syndrome,Complement regulatory protein defects
4	111	EGF	Colorectal cancer,Hypomagnesemia
4	111	LRIT3	Congenital stationary night blindness
7	7	RAC1	Autosomal dominant mental retardation
7	158	DNAJB6	Limb-girdle muscular dystrophy

Posteriormente, se calcularon las desviaciones estándar de cada ventana de las muestras colombianas con una mayor desviación estándar, (Tabla 14) y enfermedades asociadas, (Tabla 15).

Tabla 14: Ventanas de muestras colombianas con mayor desviación estándar frente a otras poblaciones

C	W	CLM	EUR	SAS	EAS	AFR	AMR
1	169	0,00197	0,0018	0,00176	0,0016	0,00175	0,00182
2	228	0,00332	0,0029	0,00271	0,00241	0,00297	0,00281
2	234	0,0087	0,007	0,00555	0,00484	0,00542	0,006
3	5	0,01467	0,0143	0,01268	0,01331	0,01248	0,01415
3	59	0,004	0,0027	0,00213	0,00131	0,00235	0,00194
3	179	0,00257	0,0023	0,00226	0,00183	0,00198	0,00227
4	122	0,00261	0,0025	0,00206	0,00156	0,00235	0,00215
4	124	0,00165	0,0015	0,00152	0,00127	0,0014	0,00154
4	160	0,00549	0,0039	0,00338	0,00368	0,00337	0,00361
6	140	0,00511	0,0044	0,00357	0,00339	0,00229	0,00404
6	146	0,00364	0,0032	0,00323	0,00292	0,00323	0,0037
6	152	0,01378	0,0104	0,01217	0,00949	0,01015	0,01295
6	158	0,00923	0,0052	0,00277	0,00152	0,00278	0,00529
7	5	0,00373	0,0033	0,00246	0,00312	0,00204	0,00317
9	140	0,00398	0,0034	0,00285	0,00258	0,0028	0,00386
11	19	0,0032	0,0011	0,00098	0,00133	0,00107	0,00238
12	67	0,01425	0,0132	0,01387	0,01297	0,00962	0,01399
13	29	0,0059	0,0018	0,00141	0,00165	0,00203	0,00099
14	33	0,0026	0,0024	0,00203	0,00189	0,00192	0,00229
15	79	0,00852	0,0083	0,00799	0,00657	0,00771	0,00832
16	86	0,00249	0,002	0,0019	0,00165	0,00209	0,00241
19	12	0,00148	0,0011	0,00099	0,00087	0,00103	0,0011
20	55	0,0144	0,0131	0,01251	0,01139	0,01224	0,01376

Tabla 15: Enfermedades relacionadas con las ventanas donde los colombianos tienen mayor variación estándar respecto a otras poblaciones

C	W	GEN	ENFERMEDAD
1	169	TBX19	Adrenocorticotropic hormone deficiency; Isolated ACDH deficiency
2	228	COL4A4	Alport syndrome,Benign familial hematuria; Thin basement membrane nephropathy

2	234	CHRND	Congenital myasthenic syndrome,Multiple pterygium síndrome
2	234	CHRNG	Multiple pterygium síndrome
2	234	DIS3L2	Perlman síndrome
2	234	ECEL1	Distal arthrogryposis
2	234	GIGYF2	Parkinson disease
2	234	KCNJ13	Leber congenital amaurosis,Snowflake vitreoretinal degeneration,Vitreoretinal degeneration
2	234	PRSS56	Microphthalmia
3	5	ITPR1	Spinocerebellar ataxia (SCA)
3	5	SUMF1	Multiple sulfatase deficiency,Sphingolipidosis
3	59	ACOX2	Congenital bile acid synthesis defect
3	59	FLNB	Atelosteogenesis type I and III,Boomerang dysplasia,Larsen syndrome,Spondylocarpotarsal synostosis syndrome
3	59	PDHB	Pyruvate dehydrogenase E1-beta deficiency,Pyruvate dehydrogenase complex deficiency
3	179	PIK3CA	Breast cancer,CLAPO syndrome,Congenital lipomatous overgrowth, vascular malformations, and epidermal nevi; CLOVE syndrome,Cowden syndrome,Cushing syndrome,Desmoplastic small round cell tumor,Hepatocellular carcinoma; Liver cancer,Megalencephaly-capillary malformation syndrome; MCAP syndrome,Ovarian cancer,Pituitary adenomas
4	122	PRDM5	Brittle cornea síndrome
4	124	BBS12	Bardet-Biedl síndrome
4	124	IL2	Graft-versus-host disease,Hashimoto thyroiditis,Type 1 diabetes mellitus
4	124	IL21	Common variable immunodeficiency,Type 1 diabetes mellitus
4	160	ETFDH	Glutaric acidemia,Secondary hyperammonemia
6	140	CITED2	Atrial septal defect,Ventricular septal defect
6	140	REPS1	Neurodegeneration with brain iron accumulation
6	146	EPM2A	Lafora disease,Progressive myoclonic epilepsy
6	152	RMND1	Combined oxidative phosphorylation deficiency
6	158	ARID1B	Autosomal dominant mental retardation,Coffin-Siris syndrome
7	5	AP5Z1	Hereditary spastic paraplegia
9	140	AGPAT2	Congenital generalized lipodystrophy
9	140	CARD9	Chronic mucocutaneous candidiasis; Familial candidiasis (CANDF)
9	140	C8G	Late complement pathway defects

9	140	INPP5E	Joubert síndrome
9	140	LHX3	Combined pituitary hormone deficiency,Growth hormone deficiency; Pituitary dwarfism
9	140	MAN1B1	Autosomal recessive mental retardation
9	140	NOTCH1	Adams-Oliver syndrome,Bicuspid aortic valve; Aortic valve disease,Breast cancer,T-cell acute lymphoblastic leukemia; T-cell acute lymphocytic leukemia
9	140	PMPCA	Autosomal recessive spinocerebellar ataxias
11	19	HPS5	Hermansky-Pudlak síndrome
11	19	LDHA	Glycogen storage disease,Glycogen storage disease type XI; Lactate dehydrogenase A deficiency,Muscle glycogen storage disease
12	67	GRIP1	Fraser síndrome
12	67	HMGA2	Uterine leiomyoma; Fibroid
13	29	CDX2	Gastric cancer
13	29	FLT3	Acute myeloid leukemia
13	29	PDX1	Maturity onset diabetes of the young (MODY),Pancreatic agenesis,Permanent neonatal diabetes mellitus
13	29	POLR1D	Treacher Collins síndrome
14	33	NUBPL	Mitochondrial complex I deficiency
15	79	CIB2	Deafness, autosomal recessive,Usher syndrome (US)
16	86	IRF8	IFN-gamma/IL-12 axis; Medelian susceptibility to mycobacterial disease (MSMD)
19	12	ACP5	Spondyloenchondrodysplasia with immune dysregulation (SPENCDI); Spondyloenchondrodysplasia (SPENCD)
19	12	CCDC151	Primary ciliary dyskinesia
19	12	DOCK6	Adams-Oliver síndrome
19	12	EPOR	Congenital polycythemia; Familial erythrocytosis (ECYT)
19	12	KANK2	Nephrotic síndrome
19	12	LDLR	Familial hypercholesterolemia; Autosomal dominant hypercholesterolaemia,Hyperlipidemia,Hyperlipoproteinemia type IIa; LDL receptor disorder
19	12	PRKCSH	Polycystic liver disease
19	12	SMARCA4	Autosomal dominant mental retardation,Coffin-Siris syndrome,Rhabdoid predisposition síndrome
20	55	MC3R	Genetic obesity

Ventanas Extremas

Finalmente, se identificaron aquellas ventanas donde los colombianos tienen el mayor o el menor valor multifractal de todas las poblaciones, obteniendo los siguientes resultados. Las ventanas con menor multifractalidad, (Tabla 16) y las enfermedades asociadas, (Tabla 17). Y las ventanas con mayor multifractalidad, (Tabla 18) y las enfermedades asociadas, (Tabla 19).

Tabla 16: Ventanas de muestras colombianas con menor MF frente a otras poblaciones

C	W	CLM	EUR	SAS	EAS	AFR	AMR
X	24	0,01465	0,01503	0,01526	0,015	0,01486	0,01488
12	112	0,00117	0,0024	0,00135	0,00424	0,00143	0,00185
19	1	0,00134	0,00161	0,0018	0,00148	0,00154	0,00154
3	15	0,00414	0,0053	0,00551	0,00527	0,00481	0,00624
3	122	0,00138	0,00152	0,00162	0,00187	0,00167	0,00148
3	173	0,00122	0,00149	0,00156	0,00149	0,00165	0,00134
4	56	0,00193	0,00212	0,00226	0,00215	0,00237	0,00212
4	104	0,0018	0,00214	0,0022	0,00206	0,00226	0,00217
7	158	0,00185	0,00218	0,00203	0,00219	0,0026	0,00217

Tabla 17: Enfermedades relacionadas con las ventanas donde los colombianos tienen menor multifractalidad respecto a otras poblaciones

C	W	GEN	ENFERMEDAD
X	24	PTCHD1	Autism; Autistic spectrum disorder; Pervasive developmental disorder
X	24	SAT1	Keratosis follicularis spinulosa decalvans
12	112	ATXN2	Spinocerebellar ataxia (SCA)
12	112	MYL2	Hypertrophic cardiomyopathy
12	112	SH2B3	Type 1 diabetes mellitus
12	112	TCTN1	Joubert síndrome
19	1	CFD	Alternative complement pathway component defects
19	1	ELANE	Neutropenic disorders
19	1	KISS1R	Central precocious puberty,Hypogonadotropic hypogonadism,Precocious puberty
3	15	CCDC174	Infantile hypotonia with psycHomotor retardation and characteristic facies
3	15	TMEM43	Arrhythmogenic right ventricular cardiomyopathy
3	15	XPC	Disorders of nucleotide excision repair,Xeroderma pigmentosum

3	122	CASR	Bartter syndrome,Calcium sensing receptor (CASR) related disease,Familial hypocalciuric hypercalcemia,Idiopathic generalized epilepsies,Neonatal hyperparathyroidism,Parathyroid carcinoma,Primary hyperparathyroidism; Familial hyperparathyroidism
3	122	ILDR1	Deafness, autosomal recessive
3	122	IQCB1	Nephronophthisis,Senior-Loken síndrome
3	173	SPATA16	Globozoospermia; Round-headed spermatozoa,Spermatogenic failure
4	56	KDR	Infantile hemangioma
4	56	KIT	Acute myeloid leukemia,Anaphylaxis,Breast cancer,Gastrotintestinal stromal tumor,Mast-cell leukemia,Piebaldism
4	56	PDGFRA	Chronic eosinophilic leukemia,Gastrotintestinal stromal tumor,Glioma,Hypereosinophilic syndrome
4	104	CISD2	Wolfram syndrome; DIDMOAD síndrome
4	104	MANBA	Glycoproteinoses,beta-Mannosidosis
4	104	NFKB1	Common variable immunodeficiency
4	104	SLC39A8	Congenital disorders of glycosylation type II
7	158	DNAJB6	Limb-girdle muscular dystrophy

Tabla 18: Ventanas de muestras colombianas con mayor MF frente a otras poblaciones

C	W	CLM	EUR	SAS	EAS	AFR	AMR
1	172	1,48242	1,48215	1,48188	1,4822	1,48155	1,48221
2	234	1,4669	1,46543	1,46371	1,4631	1,46276	1,46415
3	99	1,39389	1,39331	1,39334	1,39339	1,39236	1,39346
5	126	1,45566	1,45512	1,45521	1,45477	1,45297	1,45521
6	152	1,57844	1,57373	1,57614	1,57334	1,57339	1,57649
6	158	1,45784	1,45552	1,45522	1,45462	1,45346	1,45531
7	141	1,5449	1,54349	1,54271	1,53952	1,53973	1,54121
13	29	1,48183	1,48067	1,48093	1,481	1,4801	1,48077
14	46	1,48156	1,4814	1,48122	1,48131	1,4806	1,48128

Tabla 19: Enfermedades relacionadas con las ventanas donde los colombianos tienen mayor MF respecto a otras poblaciones

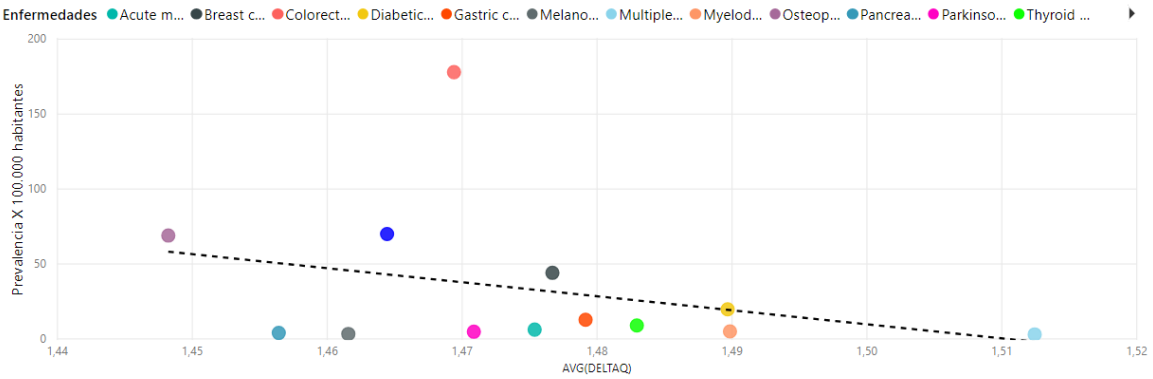
C	W	GEN	ENFERMEDAD
1	172	FMO3	Trimethylaminuria; Fish-odor síndrome
1	172	MYOC	Primary open angle glaucoma; Glaucoma 1
2	234	CHRND	Congenital myasthenic syndrome,Multiple pterygium

			syndrome
2	234	CHRNG	Multiple pterygium síndrome
2	234	DIS3L2	Perlman síndrome
2	234	ECEL1	Distal arthrogryposis
2	234	GIGYF2	Parkinson disease
2	234	KCNJ13	Leber congenital amaurosis,Snowflake vitreoretinal degeneration,Vitreoretinal degeneration
2	234	PRSS56	Microphthalmia
3	99	CPOX	Hepatic porphyria,Porphyria
5	126	ALDH7A1	Pyridoxine-dependent epilepsy
6	152	RMND1	Combined oxidative phosphorylation deficiency
6	158	ARID1B	Autosomal dominant mental retardation,Coffin-Siris syndrome
7	141	BRAF	Cardiofaciocutaneous syndrome,Langerhans cell histiocytosis,Leopard syndrome,Melanoma,Noonan syndrome,Noonan syndrome and related disorders,Thyroid cancer
13	29	CDX2	Gastric cancer
13	29	FLT3	Acute myeloid leukemia
13	29	PDX1	Maturity onset diabetes of the young (MODY),Pancreatic agenesis,Permanent neonatal diabetes mellitus
13	29	POLR1D	Treacher Collins síndrome
14	46	FANCM	Fanconi anemia,Spermatogenic failure

3.3.8 Relación entre el grado de multifractalidad y la prevalencia de enfermedades

La teoría propuesta por Moreno en [5] predice que las regiones con mayor multifractalidad son menos propensas a ciertas enfermedades, mientras las de baja multifractalidad son más sensibles a ciertas enfermedades. A fin de buscar evidencias que sugieran cierta relación entre la predicción teórica MF y las enfermedades, se buscaron las prevalencias (en los informes de salud del MSPS de Colombia y en estadísticas de la organización mundial de la salud (WHO) de las enfermedades identificadas en las ventas de baja y alta MF. En epidemiología, la prevalencia es la proporción de individuos de un grupo o una población que presentan una característica o evento determinado. La Figura 53 muestra una línea de tendencia negativa entre las dos variables, es decir a mayor prevalencia de la enfermedad menor MF de la región cromosómica, donde se localizan los genes asociados con dicha enfermedad. La muestra una línea de tendencia negativa, es decir a mayor prevalencia menor MF de la enfermedad.

Figura 53: Prevalencia de enfermedades vs Multifractalidad de la ventana donde se encuentran los genes relacionados



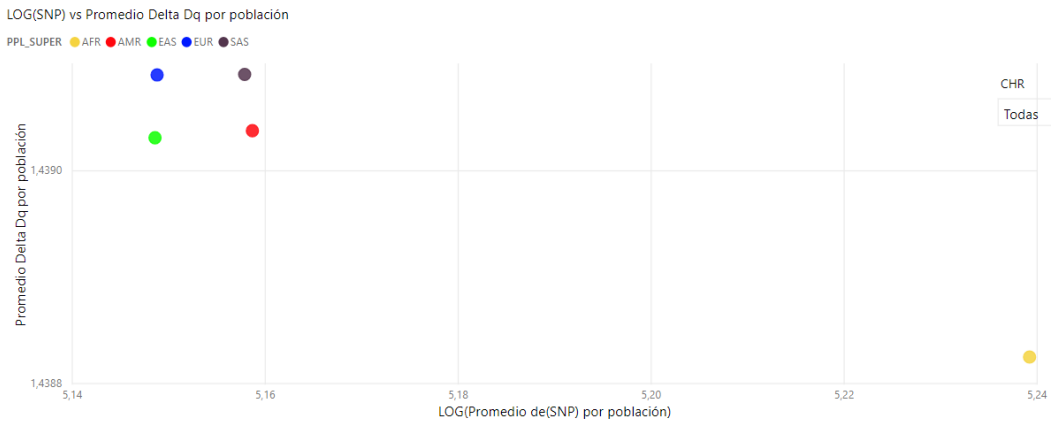
La correlación entre las dos variables no está ajustada de manera significativa debido a algunas razones. 1) La tendencia de la curva se obtuvo con sólo 14 enfermedades (Tabla 26) y en los resultados se identificaron en las ventanas más extremas para Colombia: 40 enfermedades con las MF más bajas y 36 enfermedades con las MF más altas, en total: 76 enfermedades, siendo las resaltadas en rojo las más diferenciables. En consecuencia, no existe una buena correlación ya que no se pudieron obtener las prevalencias de la mayoría de las enfermedades encontradas. 2) De las enfermedades con baja MF se identificaron 2 enfermedades cardiacas, 2 inmunológicas, 7 neurológicas, 7 tipos de cáncer y un tumor benigno, 8 metabólicas y 16 Otras (reproductivas, renal, endocrinas, distrofias, inestabilidad cromosómica, desorden sexual, etc.). 3) De las enfermedades con alta MF se identificaron 1 enfermedad cardiaca, 2 metabólicas, 4 tipos de cáncer, 2 neurológicas y otras 17 enfermedades (reproductivas, renal, endocrinas, distrofias, inestabilidad cromosómica, desorden sexual, etc.). 4) El dato del cáncer colorectal sesga mucho el análisis, ya que sin este el R2 cambiaría de 0,093 a 0,2, mejorando un poco el ajuste de la tendencia.

3.3.9 Relación entre el ΔD_q y la cantidad de SNP de colombianos vs poblaciones y superpoblaciones

Con el objeto de encontrar similitudes y diferencias las poblaciones, se utilizó una medida adicional al ΔD_q la cual fue la cantidad SNP promedio por cromosoma.

En la Figura 54 cada punto corresponde a los promedios de cada superpoblación. Se observa arriba a la izquierda a las poblaciones americanas, europeas, y medio y lejano este, abajo a la derecha se observa la superpoblación africana.

Figura 54: Valores de ΔDq vs Log(SNP) promedio para las cinco superpoblaciones.



Posteriormente se calcularon los promedios por cromosoma para cada superpoblación, población y cada individuo. Se identifican tres grupos de cromosomas según su distribución frente a la superpoblación africana (en todos ellos la cantidad de SNP de las demás superpoblaciones es menor a la superpoblación africana). El primer grupo forma triángulos donde hay al menos una superpoblación que tiene una multifractaliddad cercana a la superpoblación africana (cromosomas 5,6,7,9,10,12,15,21, X). El segundo grupo forma un cluster en la parte superior izquierda donde todas las demás superpoblaciones tienen una multifractalidad claramente mayor a la africana (cromosomas 1,2,3,4,11,13,14,18,19,22). El tercer grupo muestra los cromosomas donde la super población africana tiene una alta multifractalidad respecto a las demás superpoblaciones ubicándose en la esquina superior derecha del gráfico (cromosomas 8,16,17,20).

Figura 55: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 1

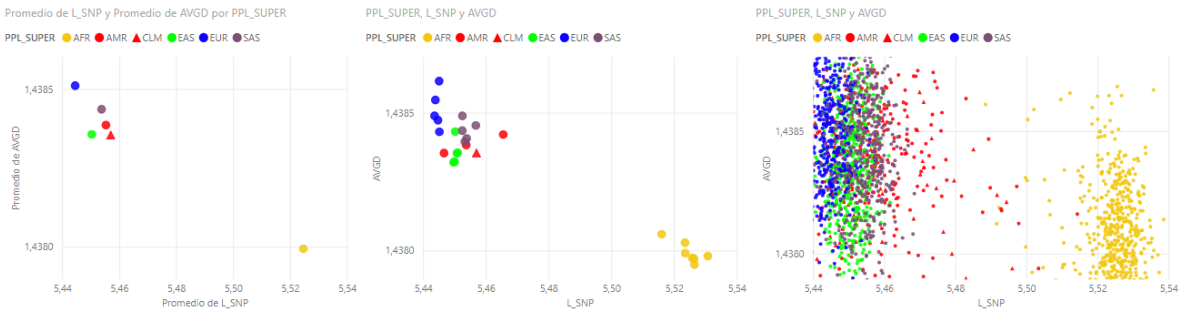


Figura 56: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 2.

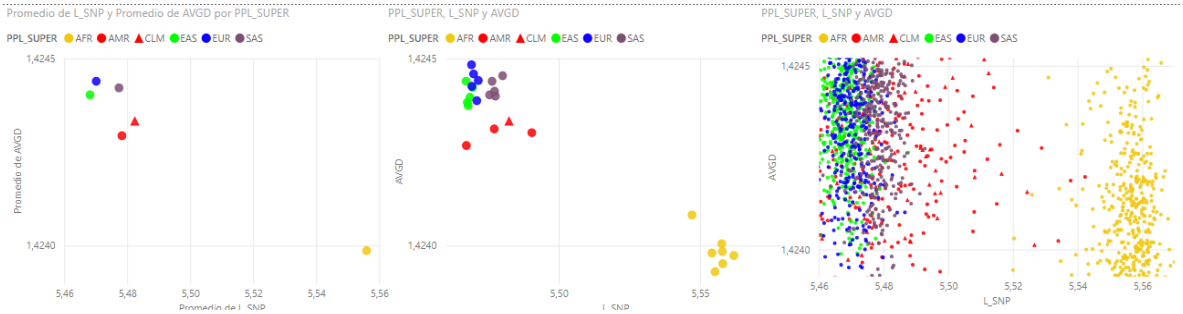


Figura 57: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 3.

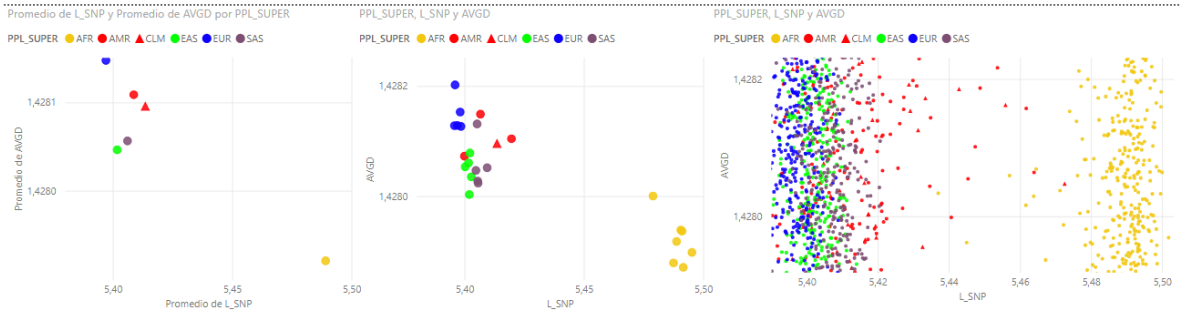


Figura 58: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 4.

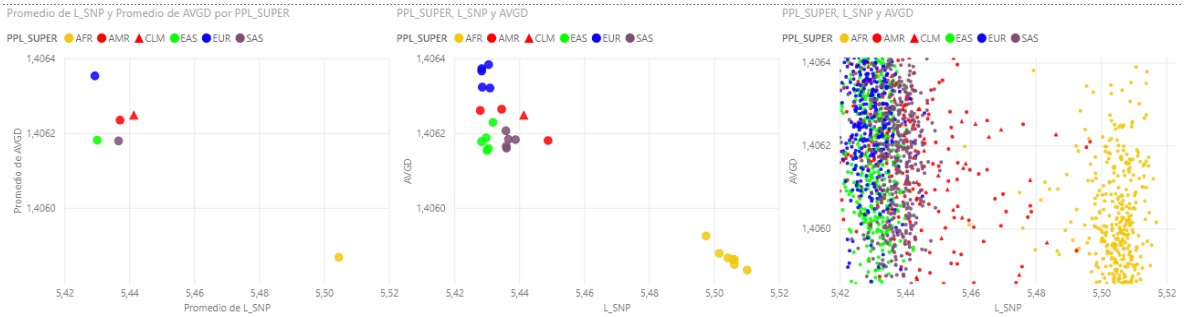


Figura 59: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 5.

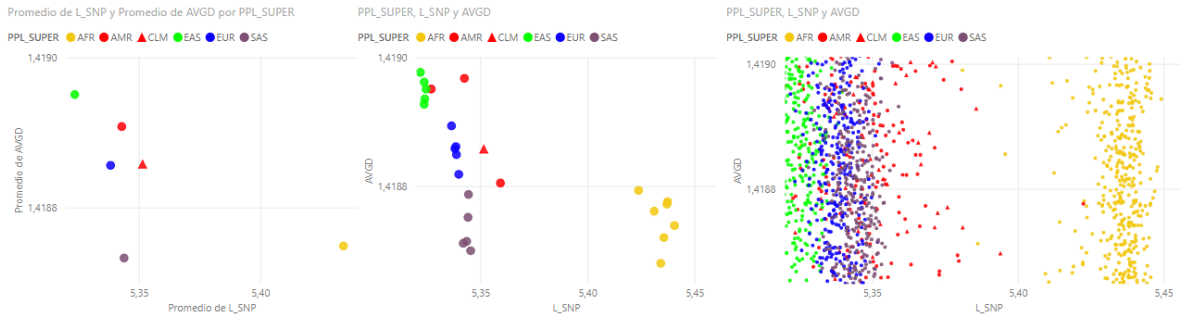


Figura 60: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 6.

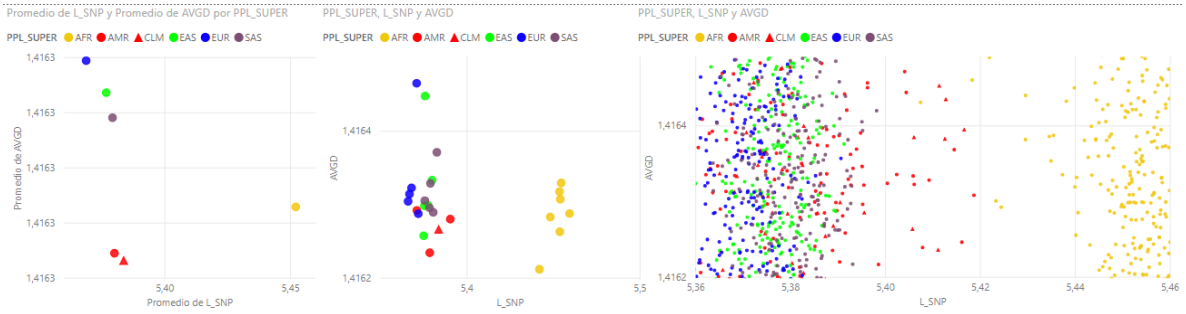


Figura 61: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 7.

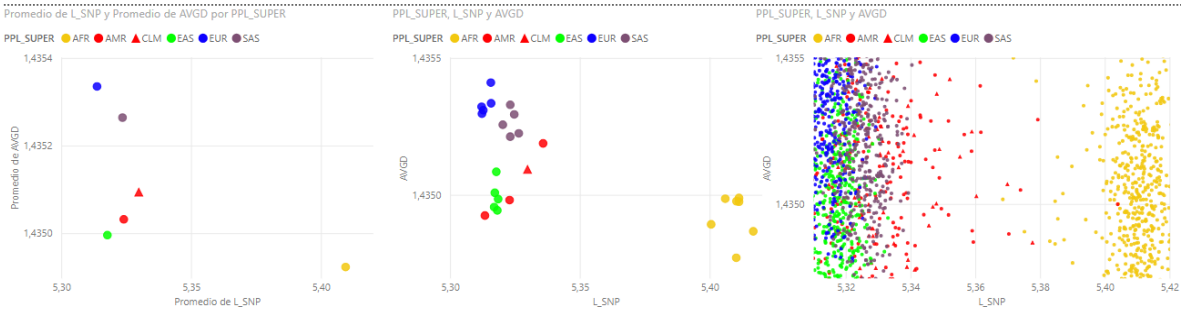


Figura 62: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 8.

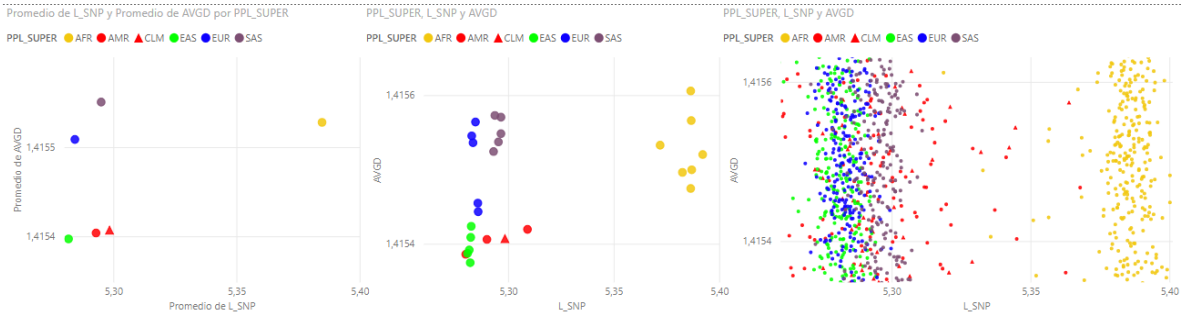


Figura 63: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 9.

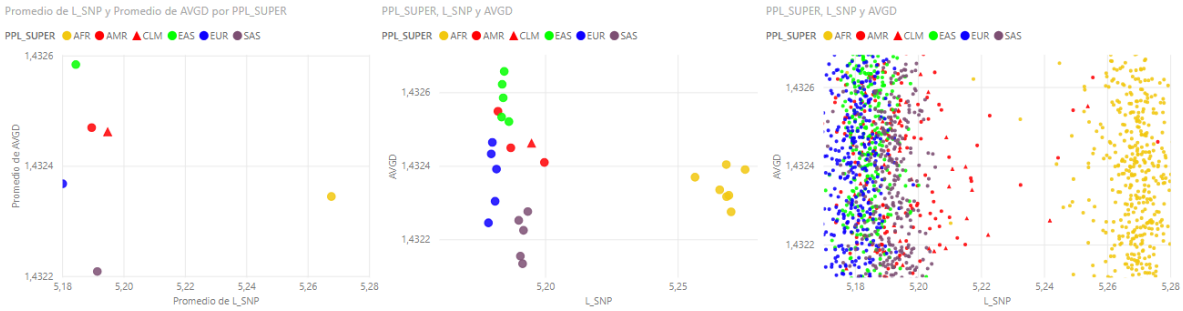


Figura 64: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 10.

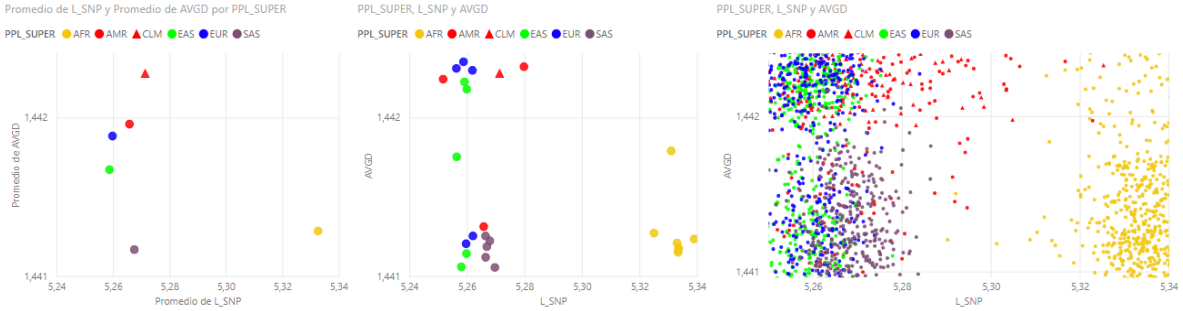


Figura 65: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 11.

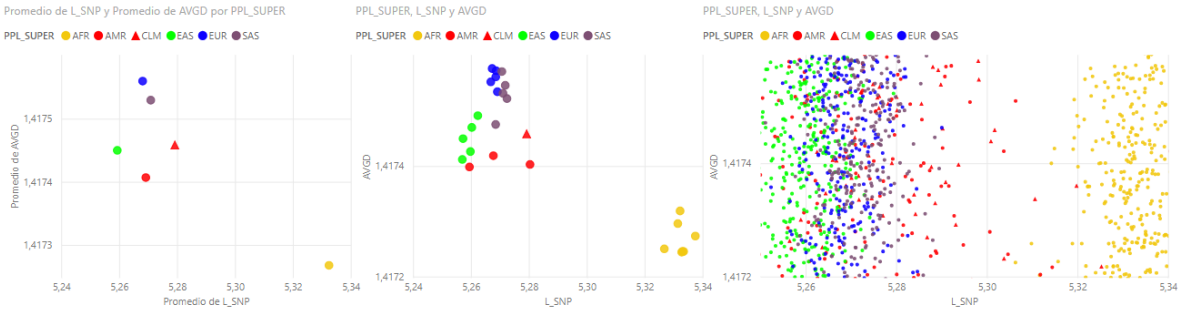


Figura 66: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 12.

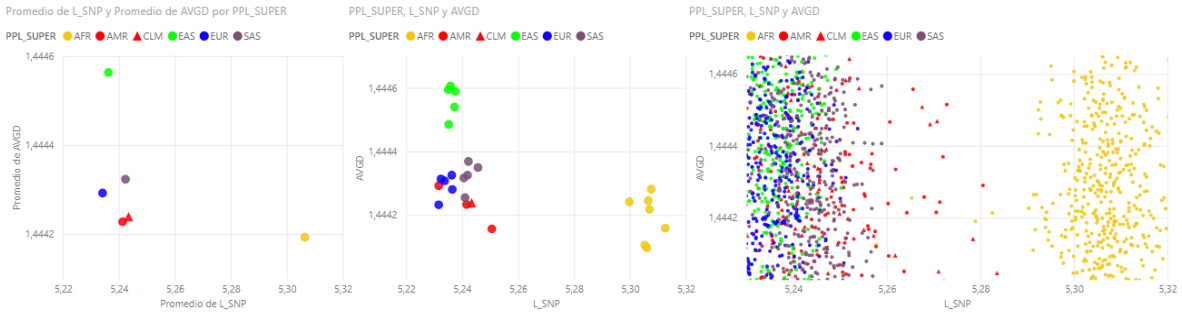


Figura 67: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 13.

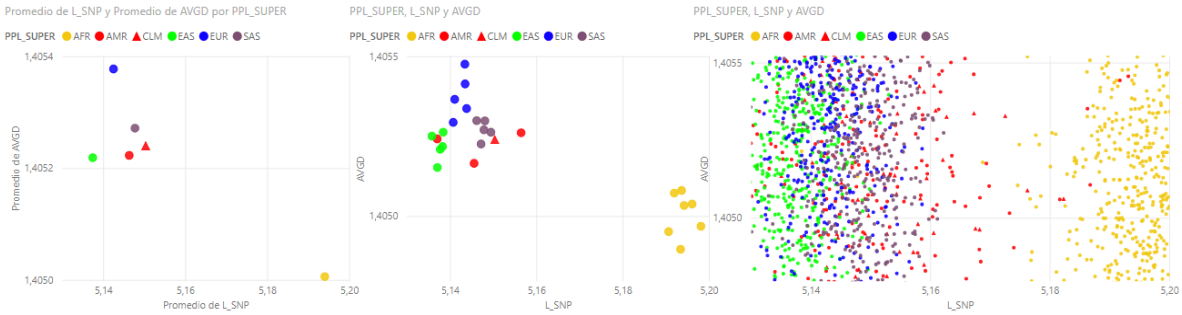


Figura 68: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 14.

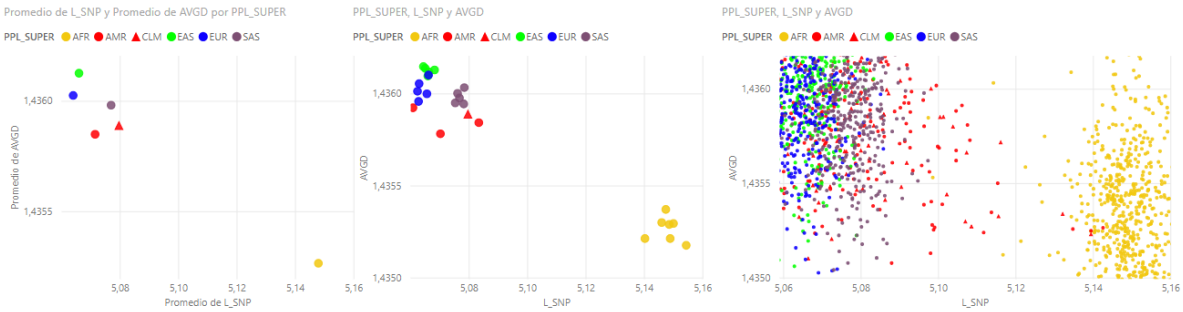


Figura 69: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 15.

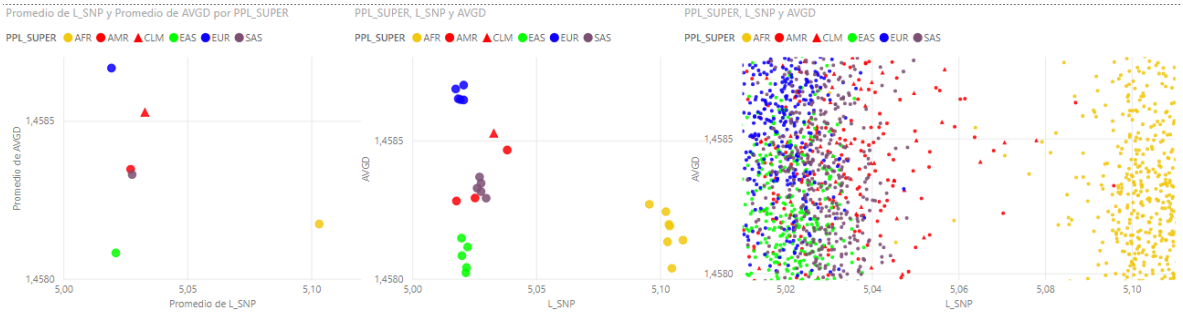


Figura 70: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 16.

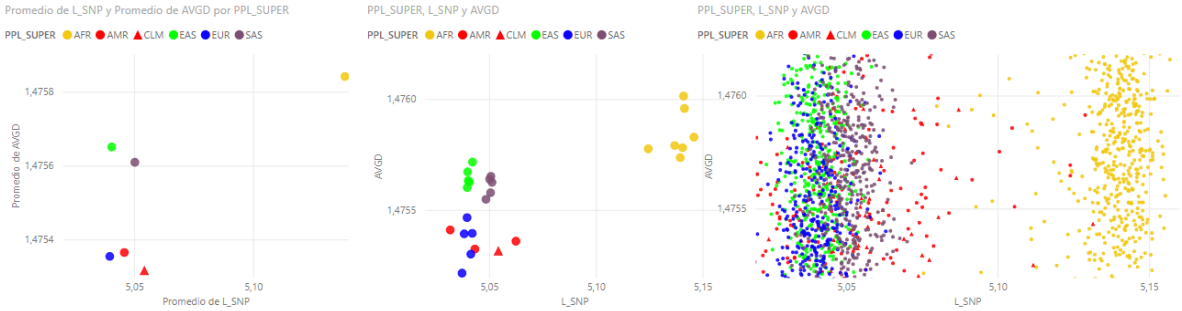


Figura 71: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 17.

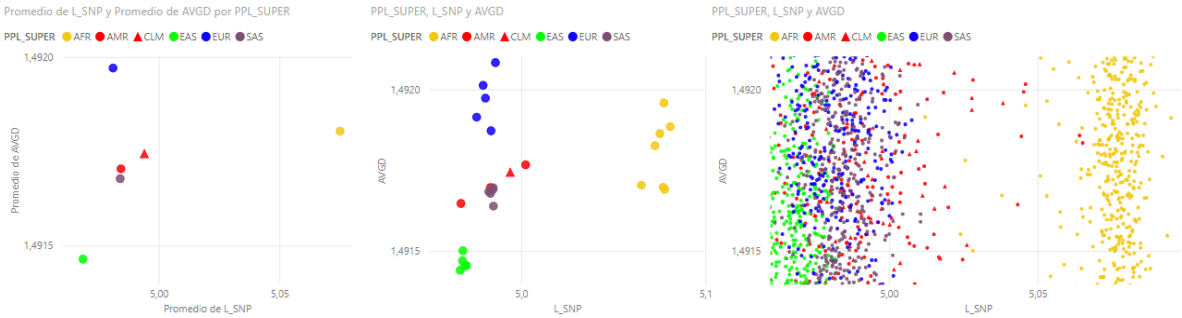


Figura 72: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 18.

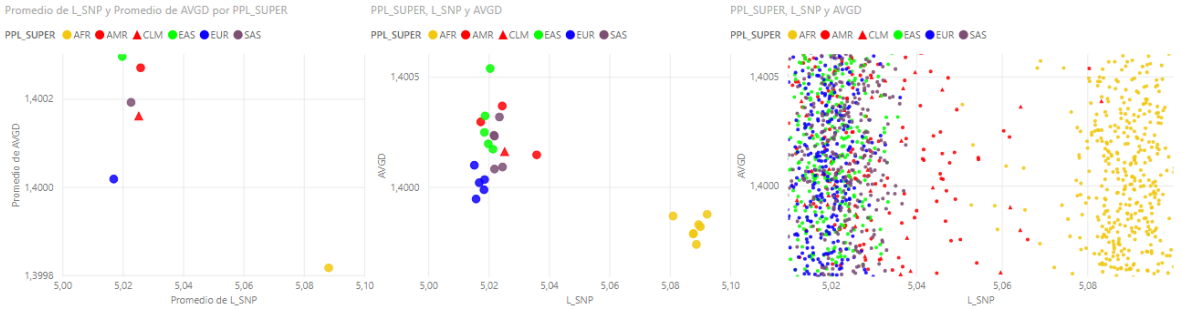


Figura 73: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 19.

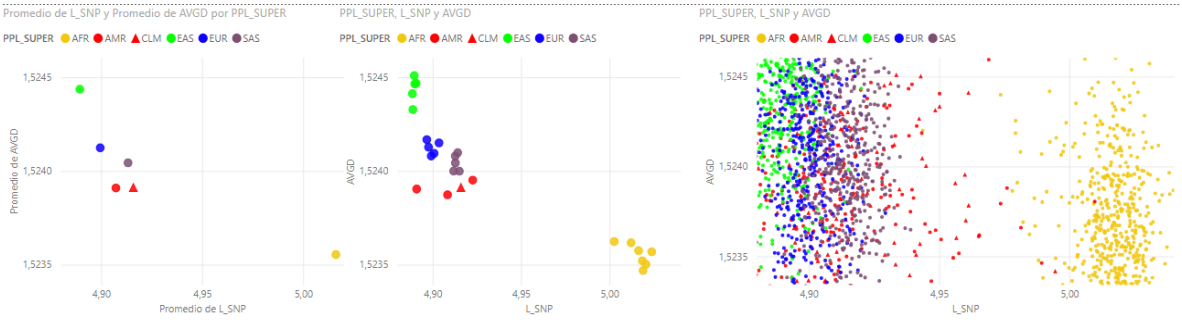


Figura 74: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 20.

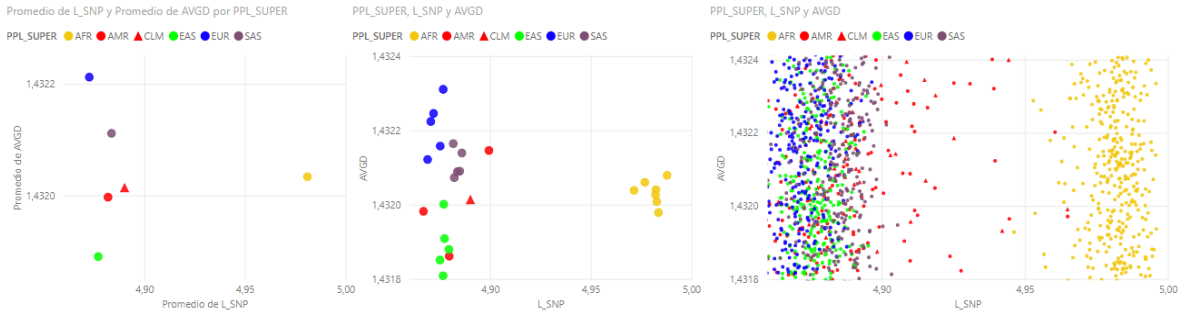


Figura 75: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 21.

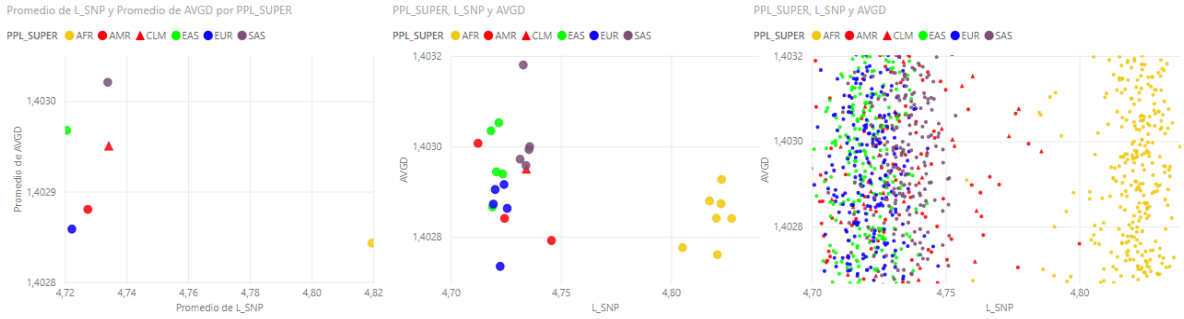


Figura 76: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma 22.

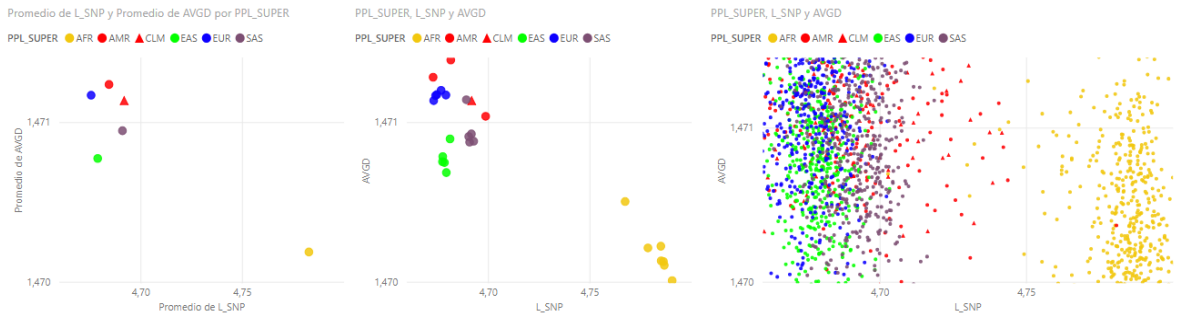
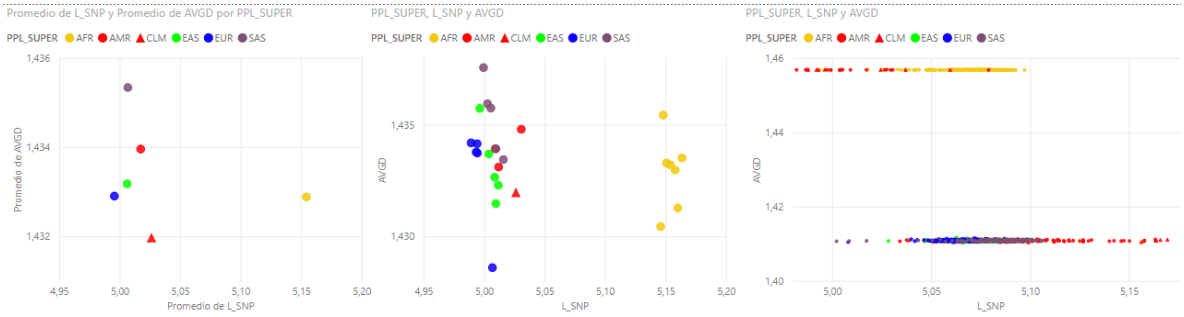


Figura 77: Valores de ΔDq vs Log(SNP) promedio por cromosoma para las cinco superpoblaciones. Cromosoma X.



3.4 Experimentos control de similitud de secuencias y multifractalidad de ventanas con alta y baja multifractalidad

Finalmente, para verificar por otros métodos la similitud y multifractalidad de algunas de estas ventanas se hizo un par de experimentos control a fin de demostrar de dos maneras diferentes la similaridad de secuencias por un lado y la presencia de baja, media y alta multifractalidad.

Por ejemplo, para determinar los grados de similaridad entre las ventanas con baja multifractalidad, media y alta multifractalidad se tomaron y analizaron las ventanas más baja y alta para una muestra colombiana versus una muestra de cada superpoblación, mediante el algoritmo de sufijo de MUMmer[92]. La Figura 78 muestra el resultado comparativo entre las secuencias.

Figura 78: Resultado de comparación de ventanas de baja multifractalidad, poca variación multifractal y alta multifractalidad, entre una muestra colombiana (referencia) y una muestra de la superpoblación EUR usando Mummer.

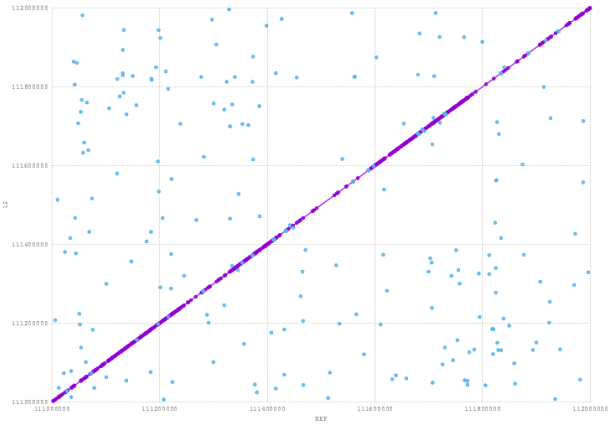
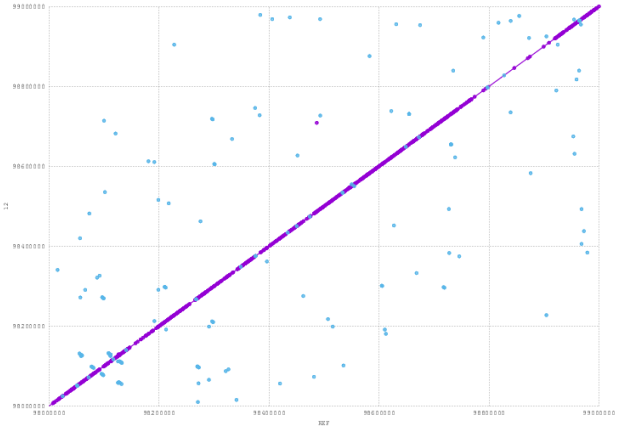
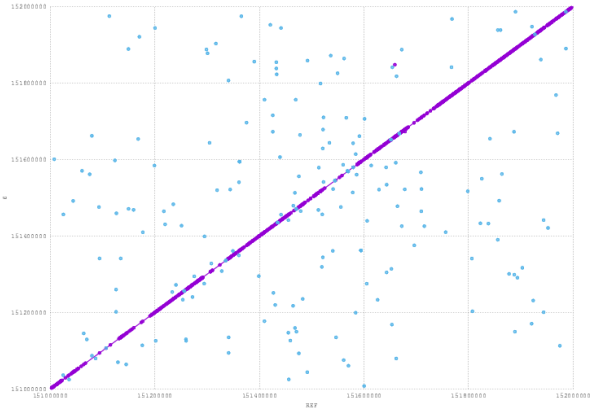
VENTANA	MUESTRA: NA12006 POBLACION: CEU (Utah) SUPERPOBLACIÓN: EUR
Baja 12-112	
Poca Variación 12-99	
Alta 6-1521	

Figura 79: Relación entre la visualización de datos y el valor multifractal en este caso para un valor bajo: Δ Dq=1,030668, correspondiente a la ventana 109 del cromosoma 14.

Por otra parte, con el objeto establecer una relación visual entre los patrones de los datos y el valor del ΔD_q se creó un mapa de calor aplicando el algoritmo de box-counting sobre el RJC de las ventanas con mayor y menor multifractalidad para un individuo colombiano, (Figura 79) y (Figura 80).

En A. se muestra un heatmap con una resolución de 32 segmentos por lado, en B. se observa la misma ventana en una resolución de 8 segmentos por lado, en C. se visualiza la misma resolución de A. pero en 3D. Los valores máximos de los picos son 0,005 y 0.004.

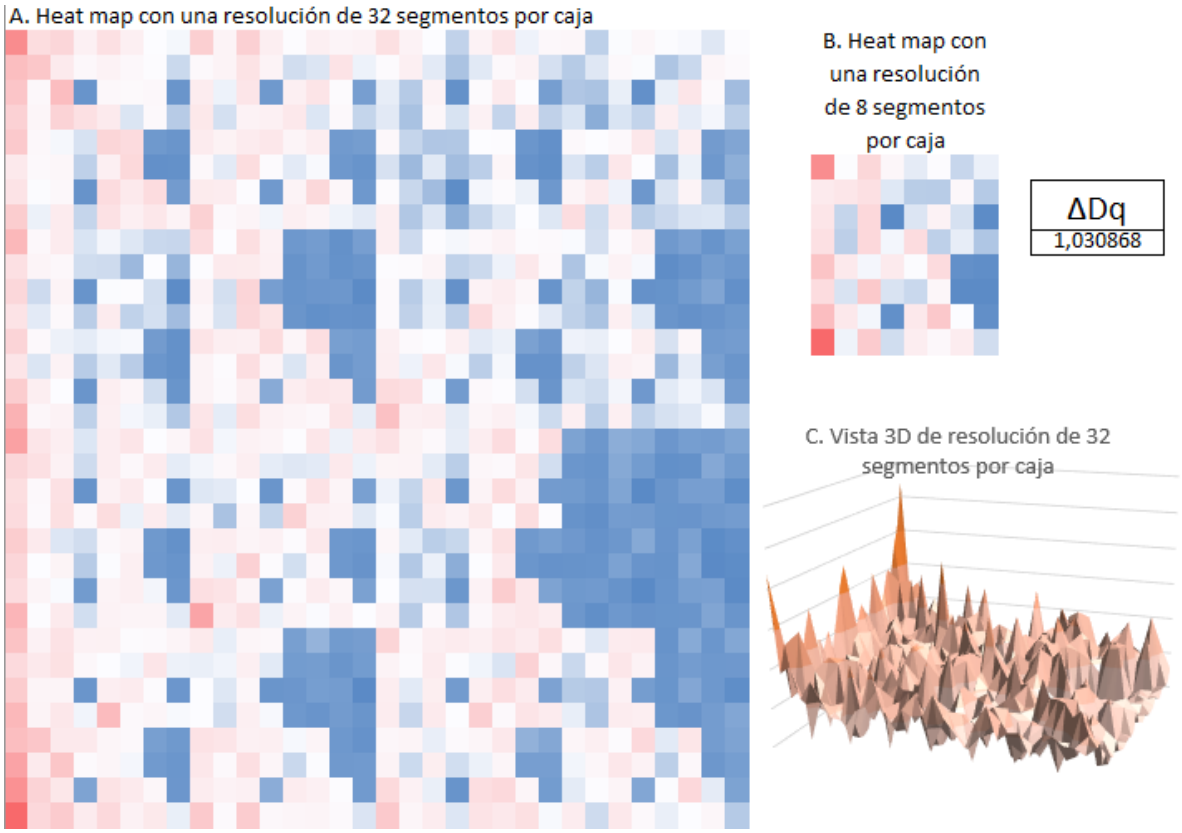
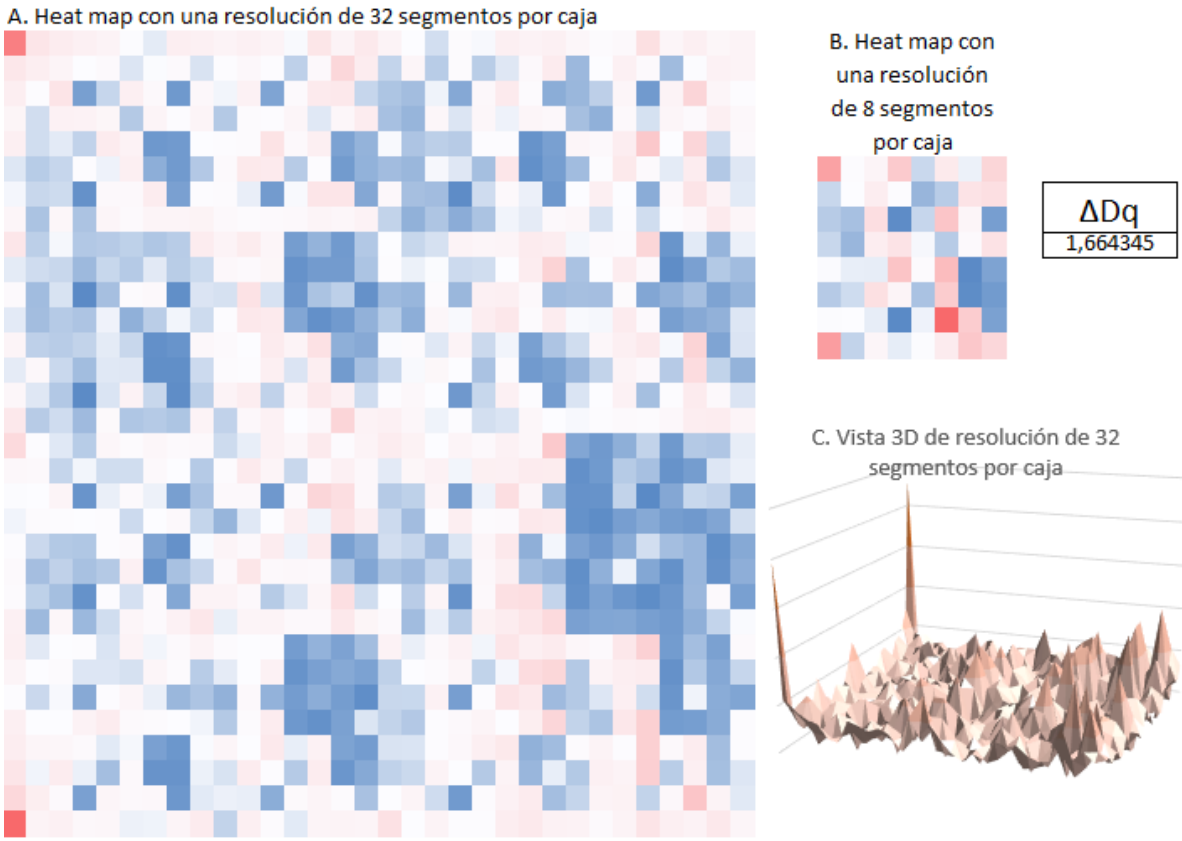


Figura 80: Relación entre la visualización de datos y el valor multifractal en este caso para un valor bajo: Δ Dq=1,664345, correspondiente a la ventana 49 del cromosoma 19.

En A. se muestra un heatmap con una resolución de 32 segmentos por lado, en B. se observa la misma ventana en una resolución de 8 segmentos por lado, en C. se visualiza la misma resolución de A. pero en 3D. Los valores máximos de los picos son 0.009 y 0.007



3.5 DISCUSION Y CONCLUSIONES

Se desarrolló de una metodología comparativa de genomas por ventanas de 1Mpb, donde se genera una representación numérica utilizando el RJC, la cual es utilizada para hallar el ΔDq mediante la técnica de conteo de cajas. Lo anterior permite efectuar una caracterización MF e identificar las zonas de especial interés por su alta o baja propensión a desarrollar enfermedades desde el punto de vista genético, teniendo en cuenta su alta o baja multifractalidad, así como la variabilidad de las ventanas. De esta manera es posible centrar esfuerzos en estudios relacionados con las demás ciencias “omicas” en las regiones más prometedoras.

La interpretación con base en las propiedades de los sistemas dinámicos no lineales o multifractales ayudan a crear un entramado no lineal para explicar las variaciones observadas en el genoma, permitiendo vislumbrar una estructura compleja y hasta ahora oculta que va más allá de patrones fijos y secuencias repetitivas monotónicas las cuales han sido constante objeto de estudio en el pasado. Se intuye una robusta arquitectura subyacente en las secuencias ADN que permite la manifestación de una gran cantidad de procesos diversos con distintos grados de sofisticación. El lograr entender su funcionamiento sería de gran importancia para la detección oportuna de riesgos y oportunidades de un organismo frente a su entorno.

Las grandes variaciones observadas en el cromosoma X ayudaron a mostrar las diferencias entre género y la prevalencia de algunas enfermedades, no solo estableció la diferencia de genero a nivel multifractal, sino que mostro ser el más variable y, en consecuencia, uno de los principales depositarios de las regiones más susceptibles a alteraciones genéticas. Dentro de la gran variedad de enfermedades relacionadas con el cromosoma X tenemos el síndrome de X frágil, distrofias musculares, hemofilia, retinosis pigmentaria e incluso infertilidad masculina.

Ancestría

Se mostró como utilizando como medidas el valor de y el número de variantes, es posible identificar clusters para los individuos de una misma población, adicionalmente, los resultados visualizados muestran claramente a Puerto Rico y Colombia como las poblaciones más cercanas a las africanas

La información combinada entre ΔDq y cantidad de SNP puede servir como medida de proximidad a nivel genético entre poblaciones humanas, es así como se observa claramente un cluster para la población africana. Las demás poblaciones también tienen sus propios agrupamientos, pero no tan marcados y aislados como los africanos (Figura 55 a la Figura 77). También se evidencia como de las demás poblaciones estudiadas, las que tenían mayor cercanía eran Puerto Rico y Colombia, confirmando así el alto porcentaje de raza negra en comparación con otras poblaciones. De lo anterior se concluye que la multifratalidad también puede ser una herramienta útil para el estudio de ancestría.

Es importante tener presente que el estudio aquí realizado se encuentra dentro del ámbito de la genética, que comprende solo una parte de los factores que determinan el desarrollo y comportamiento de un individuo incluyendo sus enfermedades, no obstante, se plantea la hipótesis de que las zonas con baja multifractalidad son más susceptibles de presentar enfermedades ya que su complejidad subyacente las hace más vulnerables a cambios y perdidas de información, por el contrario, las zonas de mayor multifractalidad deberían tender a contar con grado de resistencia mayor a los cambios bruscos y perdida de información.

Revisión de enfermedades con asociadas a las zonas de menor en la población colombiana

En este trabajo se aporta suficiente evidencia que sustenta la predicción de la teoría propuesta en [5] y que relaciona la estructura multifractal del genoma humano con la resistencia o susceptibilidad a desarrollar ciertas enfermedades. Para el caso de las muestras colombianas se determinó que son de especial interés la ventana 15 del cromosoma 3, la ventana 112 del cromosoma 12 por su aparente fragilidad y en contraste, la ventana 29 del cromosoma 13 por su mayor resistencia (MFD).

Según la teoría propuesta, estos resultados significarían que los individuos de Medellín, Colombia deberían tener una mayor prevalencia en las enfermedades cuyos genes estén localizados en regiones de bajas MF. Según, Sispro 2017 de Colombia, y las causas de morbilidad mundial de las 20 primeras causas, los colombianos en general padecen de enfermedades: cardiometabólicas, desglosadas como HTA (hipertensión arterial) que son como el ~30%, algunas enfermedades de piel, y algunas enfermedades mentales y nerviosas. En efecto, en la Tabla 17 se encuentran aquellas enfermedades que tienen la menor MF y se enmarcan en las categorías antes descritas:

- Enfermedades mentales y nerviosas: Autismo, ataxia spinocerebral, síndrome de Joubert
- Enfermedades cardiometabolicas: Cardiopatía arritmogenica, cardiopatía hipertrofica
- Enfermedades de piel: Keratosis, xerodema pigmentoso

En el anexo de material suplementario se pueden encontrar los estudios realizados sobre las enfermedades con información disponible sobre su prevalencia en Colombia y el mundo.

Representación visual de la multifractalidad

En (Figura 79 y Figura 80) se puede observar que en la ventana de menor multifractalidad , existen patrones un poco más definidos y en consecuencia observables, con picos y valles claramente agrupados y presentando y similares a diversas escalas. Por otra parte, la ventana más multifractal presenta una mayor variedad y los patrones son menos claros al presentar diversas “fractalidades” que al combinarse presentan una mayor complejidad y a su vez, mayor riqueza de información.

Un método nuevo de alineamiento MF de secuencias

Inicialmente, se desarrolló de una metodología comparativa de genomas por secciones o ventanas de longitud X, que permite identificar las zonas de especial interés por su relación y propensión a desarrollar o no enfermedades desde el punto de vista genético, teniendo en cuenta su baja o alta multifractalidad y su desviación estándar. Este método permite hacer estudios comparativos y dirigidos hacia genomas de poblaciones específicas mediante la caracterización de estas ventanas.

Actualmente, existen varios métodos para comparar secuencias largas de genomas, ya sea localmente o en línea (en la web), entre estos tenemos: MUMmer, Genome Browser, VISTA, PipMaker, NCBI Map Viewer, Ensembl, entre otros [93]. Todas estas herramientas tienen sus ventajas y desventajas. Lo significativo del enfoque generado en este trabajo es que el AMF hace parte de una teoría con elementos predictivos no lineales lo cual da ciertas ventajas frente a otros métodos tradicionales, debido a las propiedades implicadas en la teoría. Existen algunos enfoques que utilizan el AMF a partir de secuencias lineales de ADN [87], [94]. No obstante, en el presente trabajo, el enfoque utiliza la RJC para calcular los espectros MF y compararlos por ventanas o regiones cromosómicas, lo cual lo constituye en un método poderoso para estudios de genómica comparada.

La interpretación con base en las propiedades de las estructuras disipativas alejadas del equilibrio ayudan a crear un entramado "predictivo" para explicar las variaciones observadas en el genoma y su asociación con características particulares [5]. Al extrapolar las propiedades de las estructuras disipativas que se acercan o alejan del equilibrio al cromosoma, es claro que algunas regiones del cromosoma presentan características en sus secuencias que permiten predecir qué tan cerca o alejadas del equilibrio podrían estar. De hecho, la pérdida de la MF está asociada con efectos medio ambientales y cierta inestabilidad genética, y por ende, asociada con una mayor susceptibilidad (mayor prevalencia) a padecer algunas enfermedades. Por el contrario, el aumento de la MF, está relacionada con efectos deterministas y mayor estabilidad genética y, por consiguiente, asociada con menores susceptibilidades a padecer una enfermedad. En el presente trabajo las tendencias observadas son una evidencia significativa que da validez a la teoría propuesta.

Las grandes variaciones observadas en el cromosoma X fueron determinantes para diferenciar las muestras por género. Los resultados muestran que los individuos masculinos son más multifractales que las mujeres. Por otra parte, el cromosoma X, siendo aquel con menor MF en general, es el que contiene la mayor cantidad de ventanas altamente variables, y en consecuencia podría ser reflejo de la diversidad en la prevalencia de las enfermedades vinculadas a los genes ubicados en dichas ventanas. La baja multifractalidad general del cromosoma X lo convierte en uno de los principales depositarios de las regiones más susceptibles a alteraciones genéticas. Teniendo en cuenta lo anterior y el hecho de la menor MF de las mujeres se podría concluir que en general, las mujeres podrían ser más susceptibles a desarrollar enfermedades por causas genéticas, lo cual es acorde con. También se ha demostrado en otros estudios que las mujeres tienen un sistema inmune más fuerte, lo que hace que padezcan menos enfermedades infecciosas, por el contrario, ellas presentan una mayor prevalencia en enfermedades autoinmunes.

Dentro de la gran variedad de enfermedades relacionadas con el cromosoma X tenemos el síndrome de X frágil, distrofias musculares, hemofilia, retinosis pigmentaria e incluso infertilidad masculina.

Varias enfermedades localizadas en ventanas con baja MF en la población colombiana están asociadas con altas prevalencias

Según la teoría propuesta, estos resultados obtenidos de individuos de Medellín, Colombia deberían tener una mayor prevalencia en las enfermedades cuyos genes estén localizados en regiones de baja MF. A pesar de que no existe una buena correlación entre los valores de MF y las prevalencias de las enfermedades encontradas, la Figura 53 muestra una buena tendencia, la cual se podría mejorar si se pudieran obtener las prevalencias de las enfermedades faltantes. Otra alternativa para mejorar el ajuste sería progresar en el estudio de las MFs (bajas y altas) y las prevalencias de las otras poblaciones humanas presentes en el estudio (Trabajo en progreso).

Es interesante notar que en la muestra de Medellín los individuos podrían padecer principalmente de enfermedades cardiacas, inmunológicas, neurológicas, algunos tipos de cáncer (incluido el de piel) y varios tipos de enfermedades metabólicas. Según, Sispro 2017 de Colombia, y las causas de morbilidad mundial de las 20 primeras causas, los colombianos en general padecen de enfermedades: cardiometabólicas, desglosadas como HTA (hipertensión arterial) que son como el ~30%, algunas enfermedades de piel, y algunas enfermedades mentales y nerviosas. En efecto, en la Tabla 17 se encuentran aquellas enfermedades que tienen la menor MF y se enmarcan en las categorías antes descritas, entre otras:

- Enfermedades mentales y nerviosas: Autismo, ataxia spinocerebral, síndrome de Joubert.
- Enfermedades cardiometabolicas: Cardiopatía arritmogenica, cardiopatía hipertrófica.
- Enfermedades de piel: Keratosis, xerodema pigmentoso.

Enfermedades relacionadas con genes ubicados en ventanas de alta multifractalidad mostraron menor prevalencia

Se identificaron las zonas donde la población colombiana tiene una mayor MF que las demás poblaciones, sugiriendo que existe una mayor “resistencia” genética a desarrollar enfermedades relacionadas con los genes de dichos locus que otras poblaciones, muestra de ello es el cáncer de tiroides y leucemia. En este grupo también se encuentra el cáncer gástrico, sin embargo, este tiene una prevalencia mucho mayor que los dos anteriores y en el ranking mundial Colombia ocupa un puesto mayor (ver material suplementario), lo que sugiere que condiciones

medioambientales (epigenéticas) como el tipo y calidad de la dieta están presionando fuertemente el aumento de los casos.

Por otra parte, se debe tener presente que el estudio aquí realizado se encuentra dentro del ámbito de la genética y la genómica, que comprende solo una parte de los factores que determinan el desarrollo y comportamiento de un individuo y sus enfermedades. Al examinar el genoma lo que se está respondiendo es “lo que pueda pasar”. De ahí que existen otros tipos de factores epigenéticos y ambientales que pueden alterar el curso del desarrollo de una enfermedad. En ciencias ómicas es claro que el transcriptoma nos dice “lo que parece que está pasando”, la proteómica, “lo que hace que pase” y el metaboloma “lo que realmente está pasando”. De ahí la necesidad de hacer estudios de metabolómica y perfiles ómicos integrales y transversales con información clínica que permitan hacer una medicina de precisión más personalizada. A pesar de, lo más significativo del trabajo propuesto con el AMFGHC (AMF de genomas humanos completos) comparado es que con este enfoque se podría precisar de aquí en adelante donde efectuar las búsquedas y llevar acabos estudios dirigidos de perfiles ómicos. Esto permitiría ahorrar recursos y esfuerzos y concentrarse en aquellas regiones del genoma de gran interés.

La información entre el ΔDq y cantidad de SNP puede servir como medida de proximidad a nivel genético entre poblaciones humanas

Desde la Figura 55 a la Figura 77 se muestran, como la población colombiana se encuentra un poco más cerca de la africana. Las demás poblaciones (tienen también sus propios agrupamientos, pero no tan marcados y separados como los africanos. También se evidencia como de las demás poblaciones estudiadas, las más cercana a la colombiana es la de Puerto Rico, confirmando así el alto componente afro de ambas poblaciones en comparación con otras. De hecho, estos resultados están de acuerdo con lo encontrado en términos generales en un trabajo previo [King et al 2015], lo que en gran parte da validez al enfoque planteado. De lo anterior se concluye que la multifratalidad también puede ser una herramienta útil para el estudio de la ancestría.

Conclusión

Es claro que existen muchos factores que hacen que una enfermedad se desarrolle y estas no deben ser atribuibles a una causa única, sin embargo, la predisposición genética es una variable importante para tener en cuenta. Tras haber encontrado correspondencias entre las zonas de baja multifractalidad con el tipo de enfermedades de mayor prevalencia en Colombia, así como una “baja susceptibilidad” en aquellas zonas de alta multifractalidad, se fortalece la hipótesis de que mediante el análisis multifractal se pueden identificar zonas del genoma que ameriten un estudio de mayor profundidad por sus probables implicaciones sobre los grados de prevalencia de enfermedades asociadas a factores genéticos.
Se reportaron las zonas donde la población colombiana muestra una mayor multifractalidad que las demás poblaciones, sugiriendo que existe una baja susceptibilidad genética a desarrollar enfermedades relacionadas con los genes de dichos locus que otras poblaciones, muestra de ello es el cáncer de tiroides y leucemia. En este grupo también se encuentra el cáncer gástrico, sin embargo, este tiene una prevalencia mucho mayor que los dos anteriores y en el ranking mundial Colombia ocupa un puesto mayor (ver material suplementario), lo que sugiere que condiciones medioambientales (epigenéticas) como el tipo y calidad de la dieta están presionando fuertemente el aumento de los casos.

Se reportaron enfermedades vinculadas a las ventanas donde la población colombiana tiene menor multifractalidad entre las que se destacan: diabetes tipo 1, cardiopatía hipertrófica y la cardiopatica arritmogenica del ventrículo derecho, así como enfermedades de piel, lo cual se encuentra en concordancia con lo reportado por el SISPRO 2017.

BIBLIOGRAFIA

[1] S. Turajlic, A. Sottoriva, T. Graham, and C. Swanton, “Resolving genetic heterogeneity in cancer,” *Nat. Rev. Genet.*, vol. 20, no. 7, pp. 404–416, 2018.

[2] N. Acosta, R. Varela, J. A. Mesa, M. L. S. López, A. L. Cómbita, and M. C. Sanabria-Salas, “Biomarcadores de pronóstico en pacientes con cáncer de próstata localizado,” *Rev. Colomb. de Cancerol.*, vol. 21, pp. 113–125, Apr. 2017.

[3] K. Xu, Y. Shi, X. Wang, Y. Chen, L. Tang, and X. Guan, “A novel BRCA1 germline mutation promotes triple-negative breast cancer cells progression and enhances sensitivity to DNA damage agents,” *Cancer Genet.*, vol. 239, pp. 26–32, Nov. 2019.

[4] P. E. Vélez *et al.*, “The Caenorhabditis elegans genome: a multifractal analysis.,” *Genet. Mol. Res. GMR*, vol. 9, pp. 949–965, May 2010.

[5] P. A. Moreno *et al.*, “The human genome: a multifractal analysis.,” *BMC Genomic-*, vol. 12, pp. 506–506, Oct. 2011.

[6] T. Can, “Introduction to Bioinformatics,” *miRNomics: MicroRNA Biology and Computational Analysis*. Humana Press, pp. 51–71, 01-Nov-2013.

[7] J. Leipzig, “A review of bioinformatic pipeline frameworks,” *Briefings Bioinforma.*, p. 020, Mar. 2016.

[8] D. Chu, “Complexity: against systems,” *Theory Biosci.*, vol. 130, pp. 229–245, Feb. 2011.

[9] M. Gell-Mann, *The quark and the jaguar: adventures in the simple and the complex*. New York: W.H. Freeman, 2001.

[10] I. N. Herstein and D. J. Winter, *Algebra lineal y teoría de matrices*. Mexico: Grupo Editorial Iberoamérica, 1987.

[11] V. G. Ivancevic and T. T. Ivancevic, “Complex Nonlinearity Chaos, Phase Transitions, Topology Change and Path Integrals.” Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2007.

[12] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and fractals: new frontiers of science*. 2011.

[13] H. Poincaré and R. Magini, “Les méthodes nouvelles de la mécanique céleste,” *Il Nuovo Cimento*, vol. 10, pp. 128–130, Jul. 1899.

[14] A. N. Kolmogorov, “The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers,” *Proc. R. Soc. A: Math. Phys. Eng. Sci.*, vol. 434, pp. 9–13, Jul. 1991.

[15] J. T. Stuart, “Non-equilibrium Thermodynamics: Variational Techniques and Stability. Edited by R. J. DONNELLY, R. HERMAN and I. PRIGOGINE. University ofchicago Press, 1966. 313 pp. 3.60.,” *J. Fluid Mech.*, vol. 50, pp. 827–827, Dec. 1971.

[16] P. Bak, C. Tang, and K. Wiesenfeld, “Self-organized criticality: An explanation of the 1/fnoise,” *Phys. Rev. Lett.*, vol. 59, pp. 381–384, Jul. 1987.

[17] D. Fereidooni, “Seismic Hazard Assessment of the City of Khoy and Its Vicinity, NW of Iran,” *Int. J. Geotech. Earthq. Eng.*, vol. 6, pp. 15–27, Jan. 2015.

[18] “http://neocybernetics.com/report145/Chapter10.pdf.” .
[19] B. M.-C. theory and 1953, “An informational theory of the statistical structure of language,” 1953.
[20] N. Hatzigeorgiu, G. Mikros, and G. Carayannis, “Word Length, Word Frequencies and Zipf’s Law in the Greek Language,” *J. Quant. Linguist.*, vol. 8, pp. 175–185, Dec. 2001.
[21] G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann, “Towards a theory of word length distributionast,” *J. Quant. Linguist.*, vol. 1, pp. 98–106, Jan. 1994.
[22] B. Mandelbrot, *Los objetos fractales. Forma, azar y dimensión*. Tusquets editores, S.A., 1985.
[23] B. Mandelbrot, “How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension,” *Science*, pp. 636–638, 1965.
[24] “Geomería fractal y geometría euclidiana,” *Revista Educación y Pedagogía*, vol. 15. pp. 83–91, 2003.
[25] L.-Q. Zhou, Z.-G. Yu, J.-Q. Deng, V. Anh, and S.-C. Long, “A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation,” *J. Theor. Biol.*, vol. 232, pp. 559–567, Feb. 2005.
[26] T. M. Reese, A. Brzoska, D. T. Yott, D. J. Kelleher, and X.-N. Zuo, “Analyzing Self-Similar and Fractal Properties of the C. elegans Neural Network,” *PLoS ONE*, pp. 40483–40483, 2009.
[27] S. Garte, “Fractal properties of the human genome,” *J. Theor. Biol.*, vol. 230, pp. 251–260, Sep. 2004.
[28] S. Vinga and J. S. Almeida, “Universal sequence map (USM) of arbitrary discrete sequences,” *BMC Bioinforma.*, vol. 3, Feb. 2002.
[29] H. J. Jeffrey, “Chaos game representation of gene structure.,” *Nucleic acids Res.*, vol. 18, pp. 2163–2170, Apr. 1990.
[30] S. Vinga, A. M. Carvalho, A. P. Francisco, L. M. Russo, and J. S. Almeida, “Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis.,” *Algorithms Mol. Biol. AMB*, vol. 7, May 2012.
[31] C. Stan, C. P. Cristescu, and E. I. Scarlat, “Similarity analysis for DNA sequences based on chaos game representation. Case study: the albumin.,” *J. Theor. Biol.*, vol. 267, no. 4, pp. 513–518, Dec. 2010.
[32] W. Deng and Y. Luan, “Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation,” *Abstr. Appl. Anal.*, vol. 2013, pp. 1–6, 2004.
[33] D. Harte, *Multifractals*. London: Chapman & Hall, 1999.
[34] S. V. Buldyrev *et al.*, “Analysis of DNA sequences using methods of statistical physics,” *Phys. A: Stat. Mech. its Appl.*, pp. 430–438, 1990.
[35] D. Kugiumtzis and A. Provata, “Statistical analysis of gene and intergenic DNA sequences,” *Phys. A: Stat. Mech. its Appl.*, pp. 623–638, Oct. 2004.
[36] A. Bunde and S. Havlin, “Fractals in science,” *Biophys. J.*, vol. 68, pp. 391–392.
[37] N. Goldman, “Nucleotide, dinucleotide and trinucleotide frequencies explain

patterns observed in chaos game representations of DNA sequences," *Nucleic Acids Res.*, pp. 2487–2491, 1984.
[38] M. T. Swain, "Fast Comparison of Microbial Genomes Using the Chaos Games Representation for Metagenomic Applications," *Procedia Comput. Sci.*, vol. 18, pp. 1372–1381, 2011.
[39] V. V. Nair, K. Vijayan, D. P. Gopinath, and A. S. Nair, "ANN Based Classification of Unknown Genome Fragments Using Chaos Game Representation," *2010 Second. Int. Conf. Mach. Learn. Comput.*, pp. 81–85, 2002.
[40] S. M. Harmon, "Recurrence Relations, Fractals, and Chaos: Implications for Analyzing Gene Structure," Colby College, 2004.
[41] P. Deschavanne, A. Giron, J. Vilain, C. Dufraigne, and B. Fertil, "Genomic signature is preserved in short DNA fragments," *Proc. IEEE Int. Symp. Bio-Informatics Biomed. Eng.*, pp. 161–167.
[42] J. G. McNally and D. Mazza, "Fractal geometry in the nucleus," *EMBO J.*, pp. 2–3, 2007.
[43] H. Zhang, "Fractal Dimension Analysis Of Dna Sequences," *GRADCON*, 1997.
[44] A. Arneodo *et al.*, "FractalsFractal and WaveletsWavelets: What Can We Learn on Transcription and Replication from Wavelet-Based Multifractal AnalysisMultifractal analysis of DNA SequencesDNA sequence ?," Springer New York, 2009, pp. 606–636.
[45] T. A. Brown, *Genomes 3*. Garland Science, 2006.
[46] H. Pearson, "Genetics: what is a gene?," *Nature*, vol. 441, no. 7092, pp. 398–401, May 2006.
[47] E. Pennisi, "Genomics. DNA study forces rethink of what it means to be a gene.," *Sci.*, vol. 316, no. 5831, pp. 1556–1557, Jun. 2007.
[48] F. Crick, "Central dogma of molecular biology.," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970.
[49] "RefSeq https: //www.ncbi.nlm.nih.gov/refseq/about/." .
[50] L. Zhu, Y. Zhang, W. Zhang, S. Yang, J.-Q. Chen, and D. Tian, "Patterns of exon-intron architecture variation of genes in eukaryotic genomes," *BMC Genomic-*, vol. 10, p. 47, 2008.
[51] A. B. Conley *et al.*, "A Comparative Analysis of Genetic Ancestry and Admixture in the Colombian Populations of Chocó and Medellín," *G3&ampmathsemicolon#58mathsemicolon GenesvertGenomesvertGenetics*, vol. 7, pp. 3435–3447, Aug. 2017.
[52] C. D. Royal *et al.*, "Inferring Genetic Ancestry: Opportunities, Challenges, and Implications," *Am. J. Hum. Genet.*, vol. 86, pp. 661–673, May 2010.
[53] I. K. Jordan, "The Columbian Exchange as a source of adaptive introgression in human populations," *Biol. Direct*, vol. 11, Apr. 2016.
[54] A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, pp. 68–74, Aug. 2015.
[55] F.-C. Chen, T.-J. Chuang, H.-Y. Lin, and M.-K. Hsu, "The evolution of the coding exome of the Arabidopsis species - the influences of DNA methylation, relative exon position, and exon length," *BMC Evol. Biol.*, vol. 14, p. 145, 2013.

[56] D. Mouchiroud, G. D\textquotesingleOnofrio, B. Aïssani, G. Macaya, C. Gautier, and G. Bernardi, “The distribution of genes in the human genome,” *Gene*, vol. 100, pp. 181–187, Apr. 1991.
[57] S. Gelfman *et al.*, “Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons,” *Genome Res.*, vol. 22, pp. 35–50, Oct. 2011.
[58] P. A. Keane and C. Seoighe, “Intron Length Coevolution across Mammalian Genomes,” *Mol. Biol. Evol.*, vol. 33, pp. 2682–2691, Aug. 2016.
[59] D. Hollander, S. Naftelberg, G. Lev-Maor, A. R. Kornblihtt, and G. Ast, “How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing?,” *Trends Genet.*, vol. 32, pp. 596–606, Oct. 2016.
[60] W. Gilbert and M. Glynias, “On the ancient nature of introns,” *Gene*, vol. 135, pp. 137–144, Dec. 1993.
[61] L. Wang and L. D. Stein, “Modeling the evolution dynamics of exon-intron structure with a general random fragmentation process,” *BMC Evol. Biol.*, vol. 13, p. 57, 2012.
[62] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNNtextquotesingle95 - International Conference on Neural Networks*.
[63] K. J. Willis and J. C. McElwain, *The evolution of plants*. Oxford: Oxford University Press, 2013.
[64] P. Kenrick and P. R. Crane, “The origin and early evolution of plants on land,” *Nature*, vol. 389, no. 6646, pp. 33–39, 1996.
[65] C. H. Wellman, P. L. Osterloff, and U. Mohiuddin, “Fragments of the earliest land plants,” *Nature*, vol. 425, no. 6955, pp. 282–285, 2002.
[66] B. M.-B. W. and E and 1965, “Information theory and psycholinguistics,” 1965.
[67] M. Suzuki, “Phase Transition and Fractals,” *Prog. Theor. Phys.*, vol. 69, pp. 65–76, Jan. 1983.
[68] M. B. Hastings, “Fractal to Nonfractal Phase Transition in the Dielectric Breakdown Model,” *Phys. Rev. Lett.*, vol. 87, Oct. 2001.
[69] J. R. Nicolás-Carlock, J. L. Carrillo-Estrada, and V. Dossetti, “Universal fractality of morphological transitions in stochastic growth processes,” *Sci. Reports*, vol. 7, no. 1, p. 3523, 2016.
[70] A. Provata and P. Katsaloulis, “Hierarchical multifractal representation of symbolic sequences and application to human chromosomes,” *Phys. Rev.*, vol. 81, Feb. 2010.
[71] I. Prigogine, “Tan Solo Una Ilusion?,” 1993.
[72] J. W.-A. for rational mechanics and analysis and 1972, “Dissipative dynamical systems part I: General theory,” 1972.
[73] J. Sylvan and J. Huber, “Life Thrives Within the Eart’s Crust.” 01-Oct-2018.
[74] D. Y. C. Wang, S. K.-P. of the …, and 1999, “Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi,” 1999.
[75] “Wikipedia: Explosión del cambrico.” .
[76] I. G. Gass, P. J. Smith, and R. C. L. Wilson, “Introducción a las ciencias de la tierra,” 1978.

[77] J. D. Hawkin, “A survey on intron and exon lengths,” *Nucleic Acids Res.*, vol. 16, pp. 9893–9908, 1987.
[78] L. Fedorova and A. Fedorov, “Introns in gene evolution.,” *Genetica*, vol. 118, no. 2–3, pp. 123–131, Jul. 2003.
[79] I. B. Rogozin, L. Carmel, M. Csuros, and E. V. Koonin, “Origin and evolution of spliceosomal introns,” *Biol. Direct*, vol. 7, p. 11, 2011.
[80] A. Anna and G. Monika, “Splicing mutations in human genetic disorders: examples, detection, and confirmation,” *J. Appl. Genet.*, vol. 59, pp. 253–268, Apr. 2018.
[81] J.-M. Claverie, “From Bioinformatics to Computational Biology,” *Genome Res.*, vol. 10, pp. 1277–1279, Sep. 2000.
[82] F. Tobar-Tosse, P. E. Veléz, E. Ocampo-Toro, and P. A. Moreno, “Structure, clustering and functional insights of repeats configurations in the upstream promoter region of the human coding genes,” *BMC Genomic-*, vol. 19, no. 8, p. 862, 2017.
[83] T. A. Manolio, “Genomewide Association Studies and Assessment of the Risk of Disease,” *New Engl. J. Med.*, vol. 363, pp. 166–176, Jul. 2010.
[84] “chocogen.com.” .
[85] S. Q. Wong *et al.*, “Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours,” *Sci. Reports*, vol. 3, pp. 3494 EP-, Dec. 2013.
[86] Z. G. Yu, V. Anh, and K. S. Lau, “Measure representation and multifractal analysis of complete genomes.,” *Phys Rev Stat Nonlin Soft Matter Phys*, vol. 64, no. 3 Pt 1, p. 031903, Sep. 2001.
[87] C. Beck and A. Provata, “Multifractal information production of the human genome,” *EPL*, vol. 95, p. 58002, Aug. 2011.
[88] S. Tetsuka, “Difficulty in determining the association of a single nucleotide polymorphism in the ZNF512B gene with the risk and prognosis of amyotrophic lateral sclerosis,” *Rinsho Shinkeigaku*, vol. 57, pp. 417–424, 2016.
[89] T.-L. Yang, R.-H. Hao, Y. Guo, C. J. Papasian, and H.-W. Deng, “Chapter 4 - Copy Number Variation,” in *Genetics of Bone Biology and Skeletal Disease (Second Edition)*, Academic Press, 2017, pp. 43–54.
[90] P. C. Ivanov *et al.*, “Multifractality in human heartbeat dynamics,” *Nature*, vol. 399, no. 6735, pp. 461–465, 1998.
[91] B. M.- Word and 1954, “Structure Formelle des Textes et Communication: Deux Études Par,” 1954.
[92] S. Kurtz *et al.*, “Versatile and open software for comparing large genomes,” *Genome Biol.*, vol. 5, p. 12, 2003.
[93] L. A. Pennacchio and E. M. Rubin, “Comparative genomic tools and databases: providing insights into the human genome,” *J. Clin. Investig.*, vol. 111, pp. 1099–1106, Apr. 2003.
[94] Z.-G. Yu, V. Anh, and K.-S. Lau, “Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome,” *Phys. A: Stat. Mech. its Appl.*, vol. 301, pp. 351–361, Dec. 2001.
[95] B. J. Maron, M. S. Maron, and C. Semsarian, “Genetics of hypertrophic

cardiomyopathy after 20 years: clinical perspectives.," *J. Am. Coll. Cardiol.*, vol. 60, no. 8, pp. 705–715, Aug. 2012.
[96] "Genética molecular de las miocardiopatías," *Revista Española de Cardiología*, vol. 55. pp. 292–302, 01-Mar-2002.
[97] "Diabetes prevalence (% of population ages 20 to 79) - Country Ranking." .
[98] "Cancer.net - melanoma." .
[99] J. J. Lee *et al.*, "Targeted next-generation sequencing reveals high frequency of mutations in epigenetic regulators across treatment-naïve patient melanomas.," *Clin. epigenetics*, vol. 7, p. 59, Jun. 2015.
[100] "CANCER TODAY - melanoma." .
[101] J.-F. ZHANG, L.-S. QU, X.-F. QIAN, B.-L. XIA, Z.-B. MAO, and W.-C. CHEN, "Nuclear transcription factor CDX2 inhibits gastric cancer-cell growth and reverses epithelial-to-mesenchymal transition in vitro and in vivo," *Mol. Med. Reports*, vol. 12, pp. 5231–5238, Jul. 2015.
[102] "Mortality – ASR (World) vs Incidence – ASR (World), stomach, in 2018, both sexes, all ages." .
[103] "Cancer.net - Leucemia." .
[104] "CANCER TODAY - estadisticas leucemia 2018." .
[105] "CANCER TODAY - Estadisticas canger de tiroides." .

MATERIAL SUPLEMENTARIO

Tabla 20: Cantidad de ventanas por cromosoma con mayor desviación estándar

Cromosoma	Mayores al 2%	Mayores al 1.5%	Mayores al 1%
1	1	7	20
2	0	3	22
3	0	2	10
4	0	2	9
5	0	1	11
6	0	3	17
7	0	2	13
8	0	1	10
9	0	2	12
10	0	3	13
11	0	0	8
12	0	2	12
13	0	1	8
14	1	1	7
15	0	0	8
16	0	0	11
17	0	1	7
18	1	5	18
19	0	0	3
20	0	0	3
21	0	0	3
22	0	1	1
X	102	128	132
TOTAL	**105**	**165**	**358**

Gráficos de relación entre cantidad de SNP y ΔDq por población por cromosoma.

Figura 81: Valores de Δ Dq vs Log(SNP) del cromosoma 1 por población. Colombia representado con un triángulo rojo.

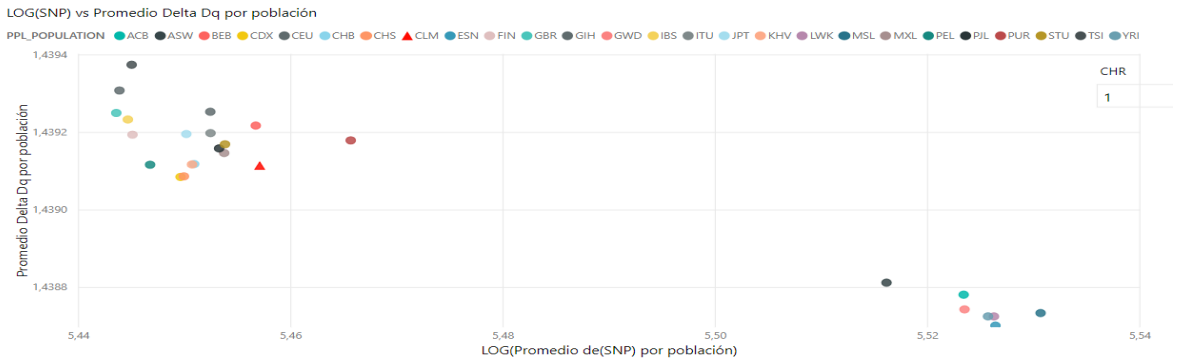


Figura 82: Valores de Δ Dq vs Log(SNP) del cromosoma 2 por población. Colombia representado con un triángulo rojo.

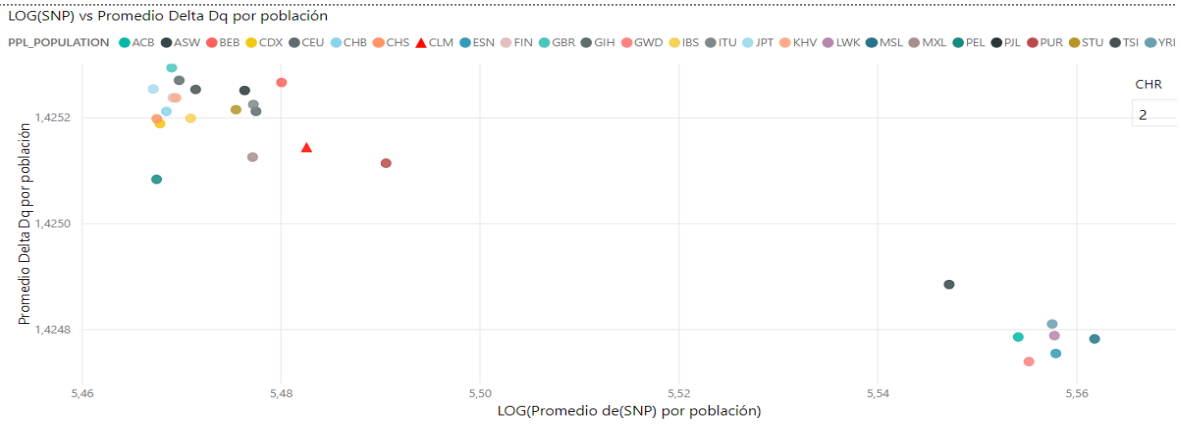


Figura 83: Valores de Δ Dq vs Log(SNP) del cromosoma 3 por población. Colombia representado con un triángulo rojo.

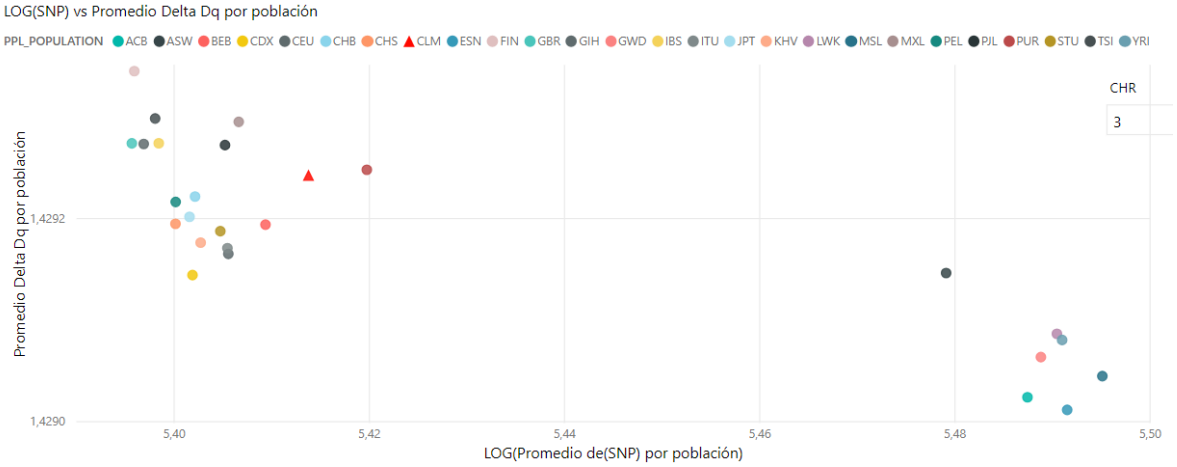


Figura 84: Valores de Δ Dq vs Log(SNP) del cromosoma 4 por población. Colombia representado con un triángulo rojo.

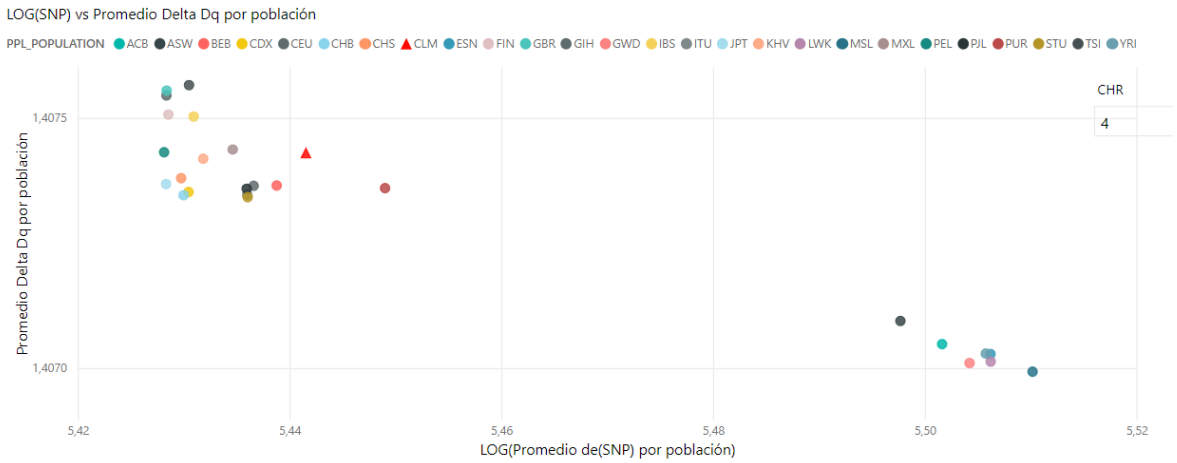


Figura 85: Valores de Δ Dq vs Log(SNP) del cromosoma 5 por población. Colombia representado con un triángulo rojo.

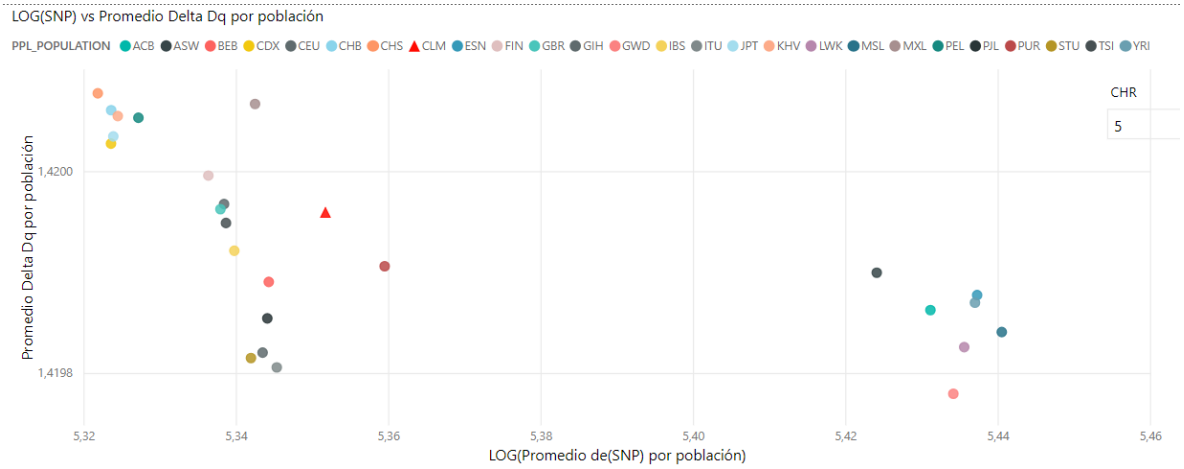


Figura 86: Valores de ΔDq vs Log(SNP) del cromosoma 6 por población. Colombia representado con un triángulo rojo.

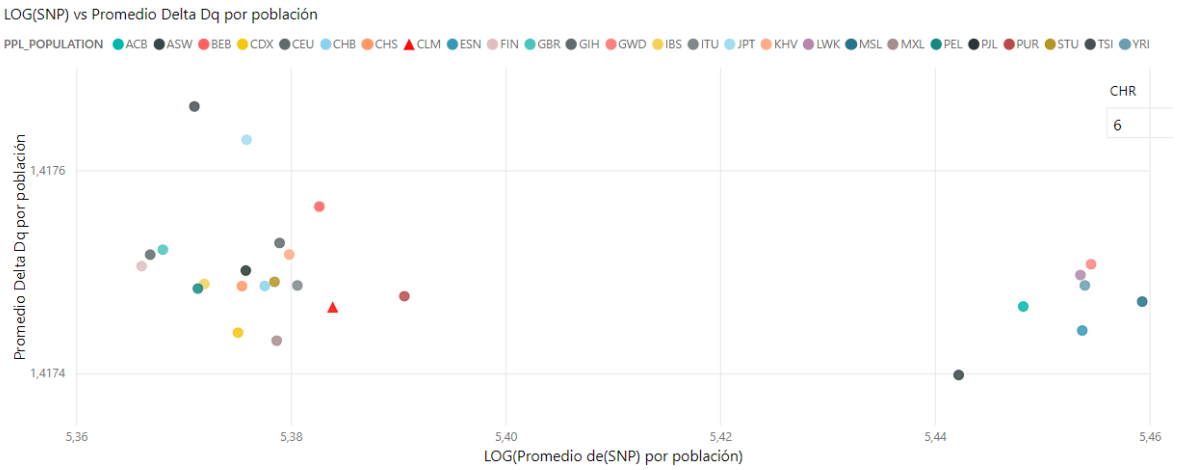


Figura 87: Valores de ΔDq vs Log(SNP) del cromosoma 7 por p oblación. Colombia representado con un triángulo rojo.

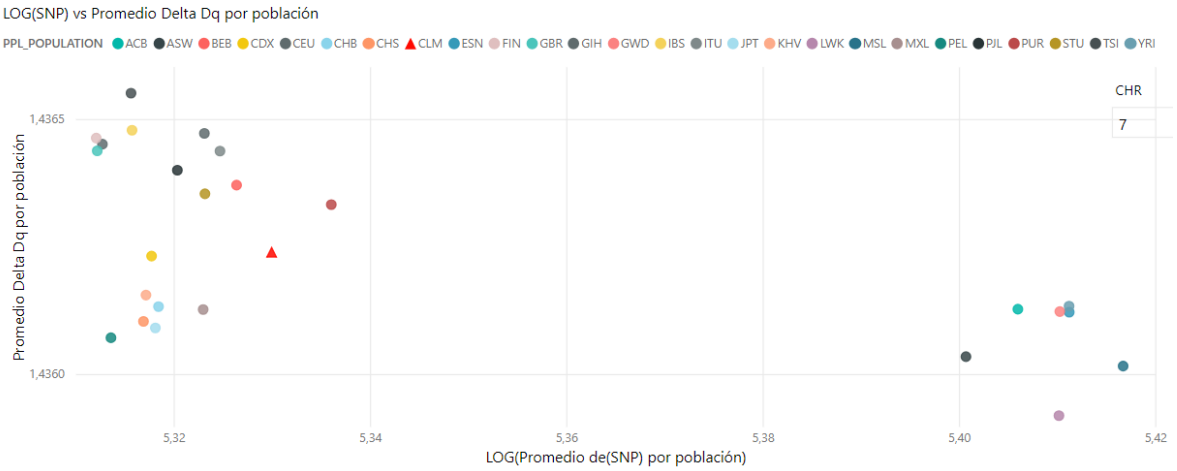


Figura 88: Valores de Δ Dq vs Log(SNP) del cromosoma 8 por población. Colombia representado con un triángulo rojo.

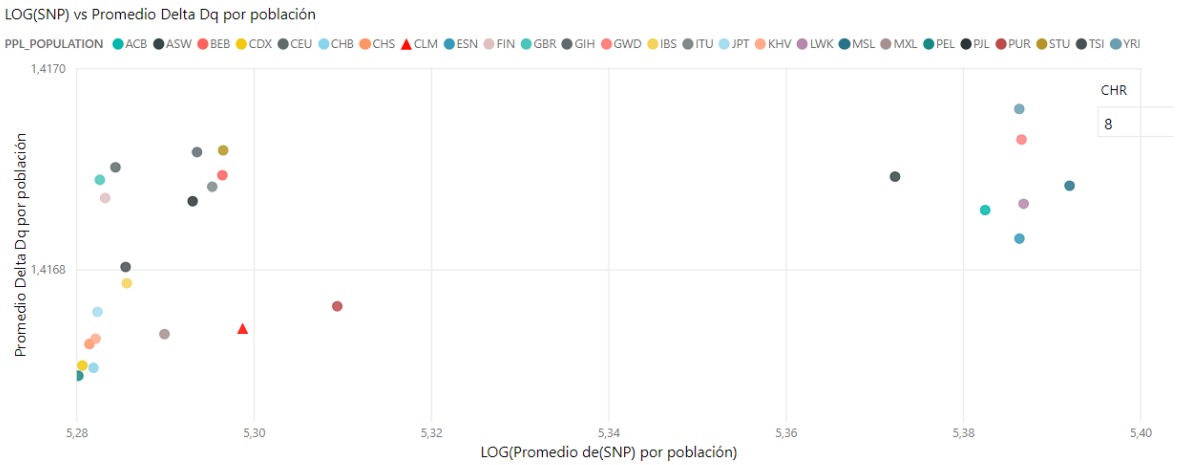


Figura 89: Valores de Δ Dq vs Log(SNP) del cromosoma 9 por población. Colombia representado con un triángulo rojo.

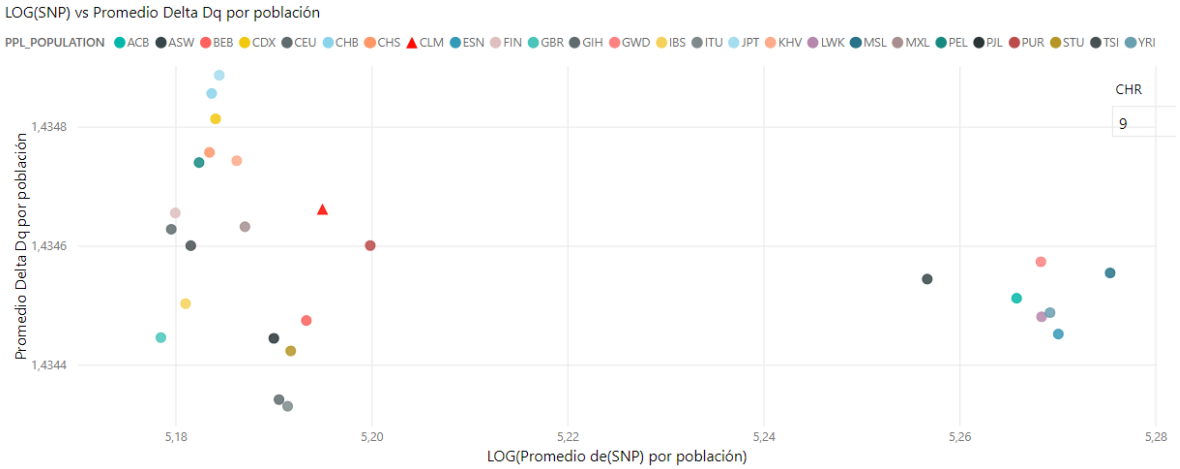


Figura 90: Valores de Δ Dq vs Log(SNP) del cromosoma 10 por población. Colombia representado con un triángulo rojo.

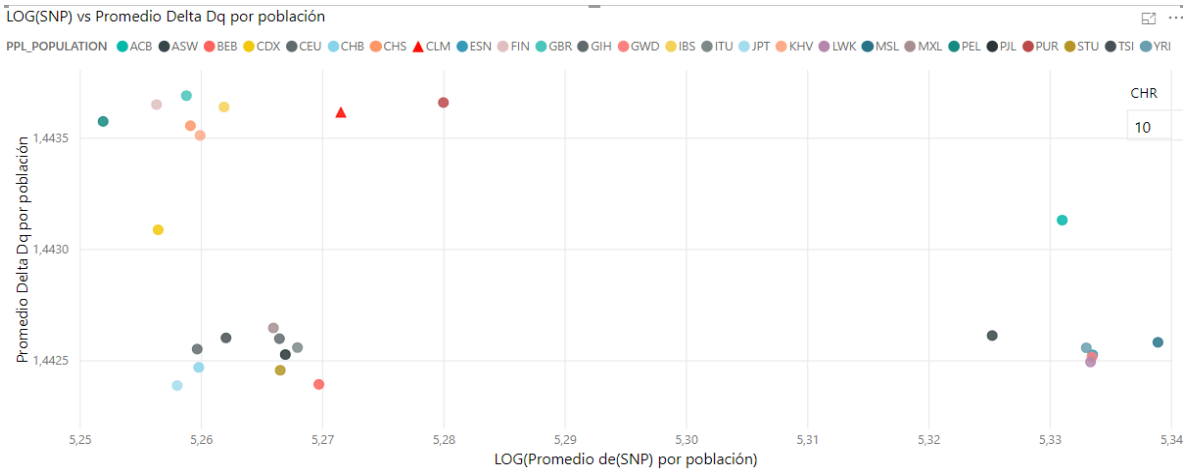


Figura 91: Valores de Δ Dq vs Log(SNP) del cromosoma 11 por población. Colombia representado con un triángulo rojo.

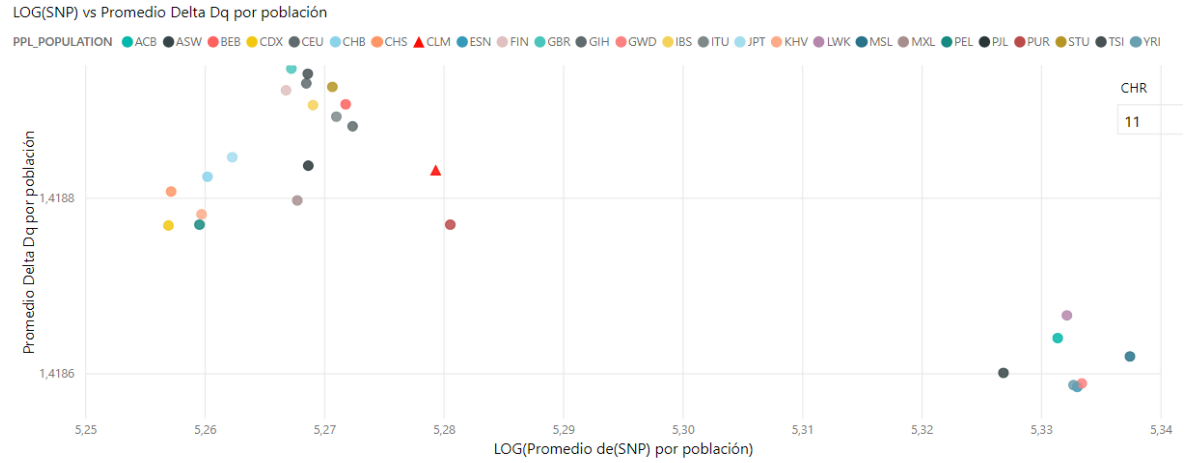


Figura 92: Valores de Δ Dq vs Log(SNP) del cromosoma 12 por población. Colombia representado con un triángulo rojo.

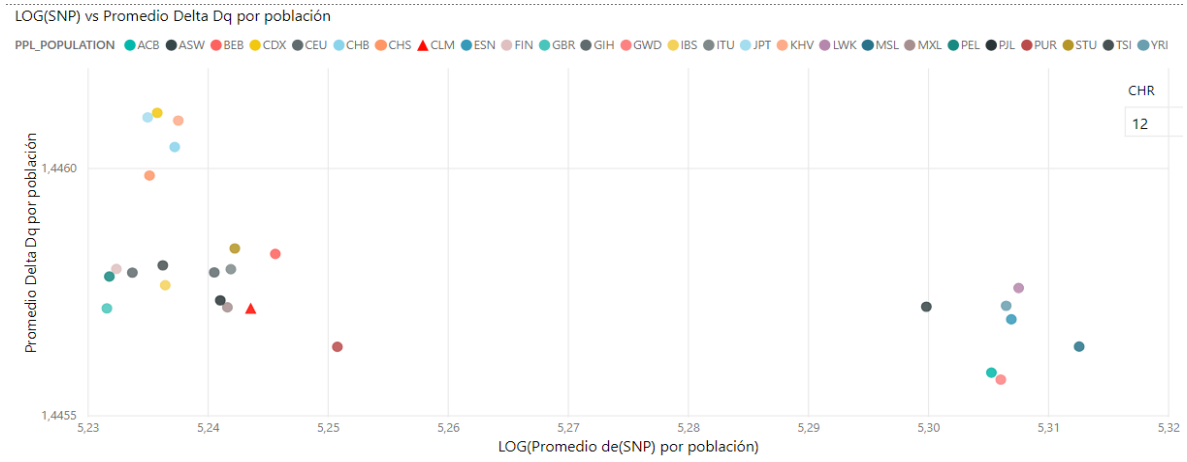


Figura 93: Valores de Δ Dq vs Log(SNP) del cromosoma 13 por población. Colombia representado con un triángulo rojo.

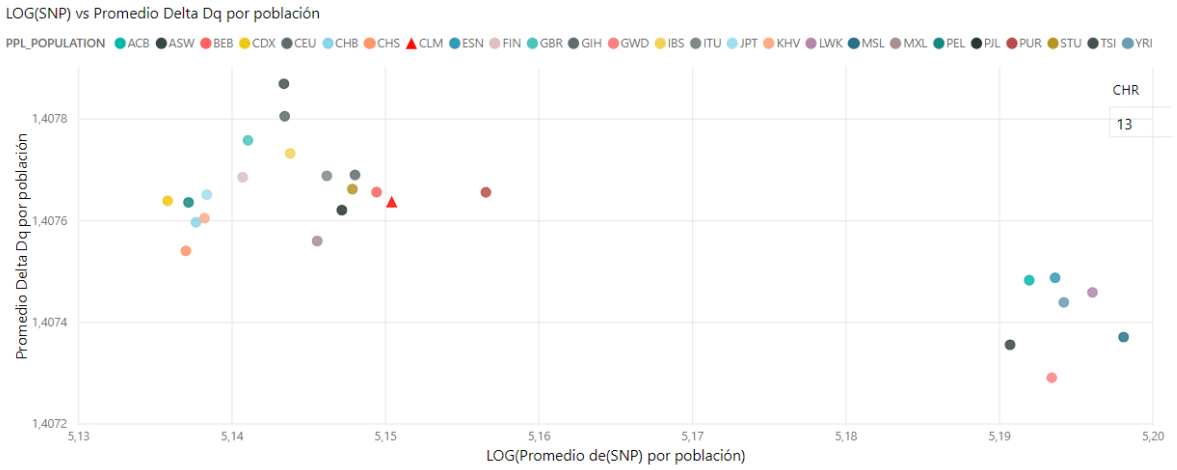


Figura 94: Valores de Δ Dq vs Log(SNP) del cromosoma 14 por población. Colombia representado con un triángulo rojo.

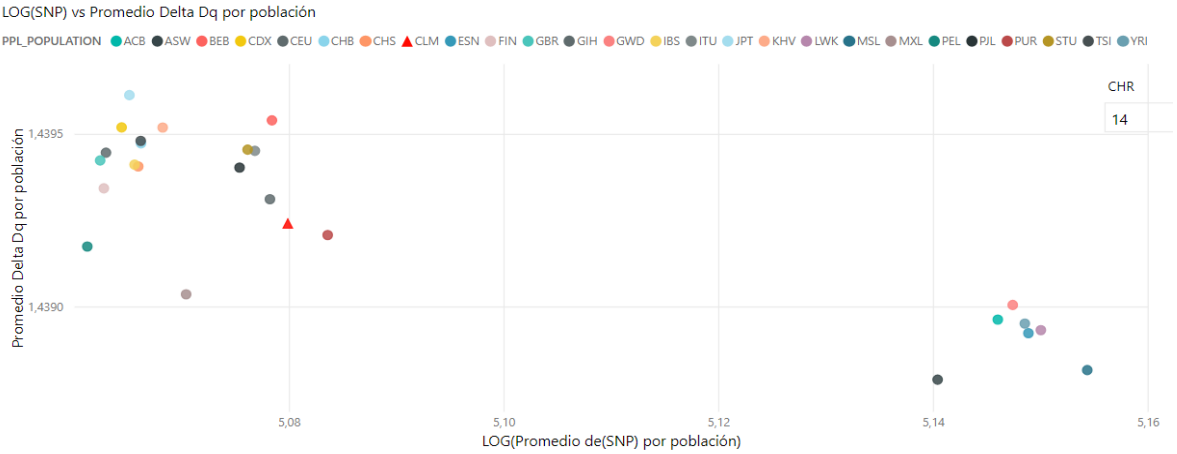


Figura 95: Valores de Δ Dq vs Log(SNP) del cromosoma 15 por población. Colombia representado con un triángulo rojo.

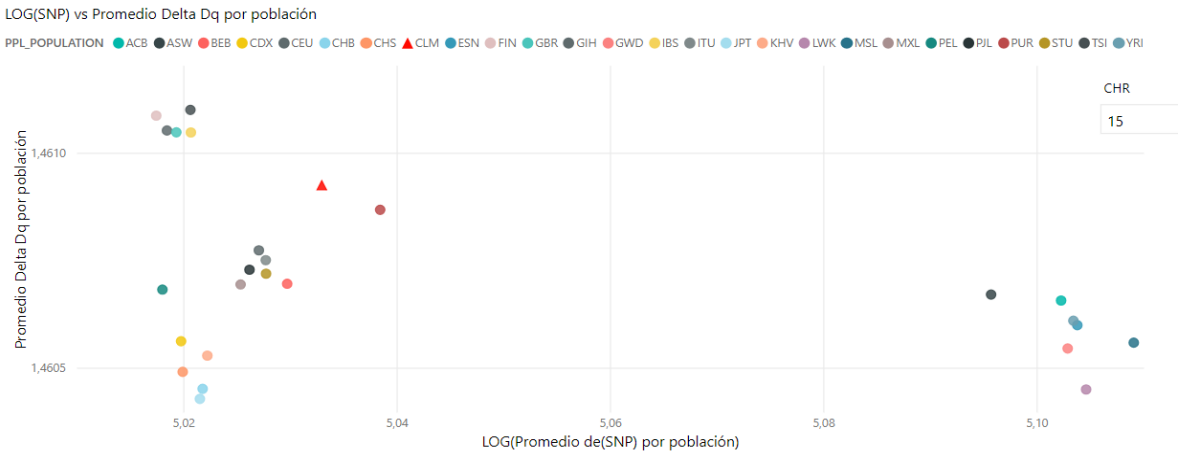


Figura 96: Valores de Δ Dq vs Log(SNP) del cromosoma 16 por población. Colombia representado con un triángulo rojo.

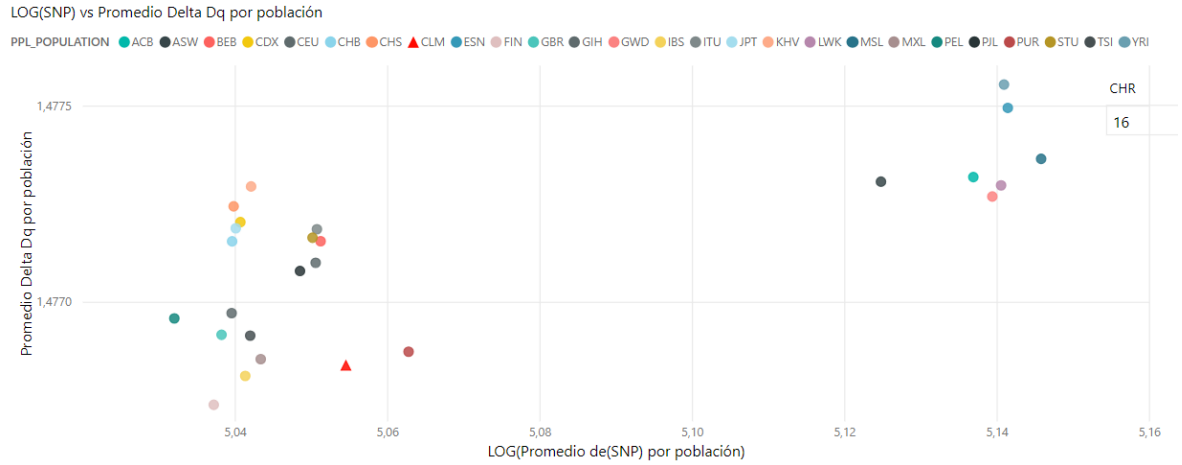


Figura 97: Valores de Δ Dq vs Log(SNP) del cromosoma 17 por población. Colombia representado con un triángulo rojo.

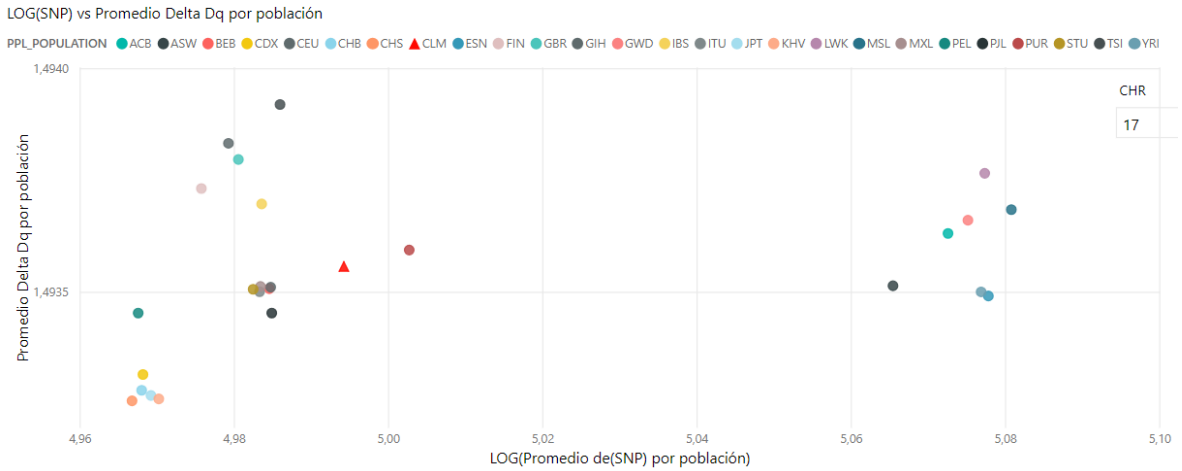


Figura 98: Valores de Δ Dq vs Log(SNP) del cromosoma 18 por población. Colombia representado con un triángulo rojo.

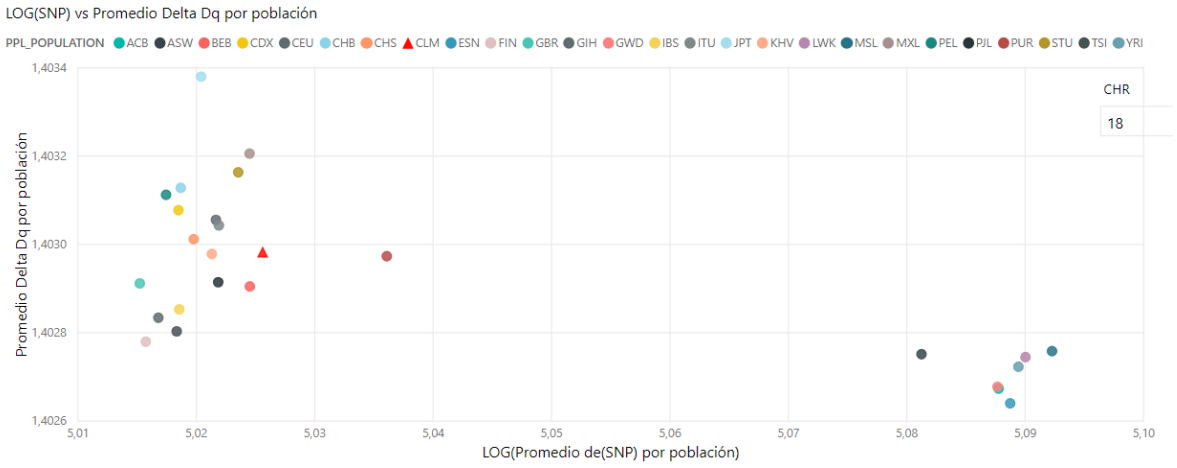


Figura 99: Valores de Δ Dq vs Log(SNP) del cromosoma 19 por población. Colombia representado con un triángulo rojo.

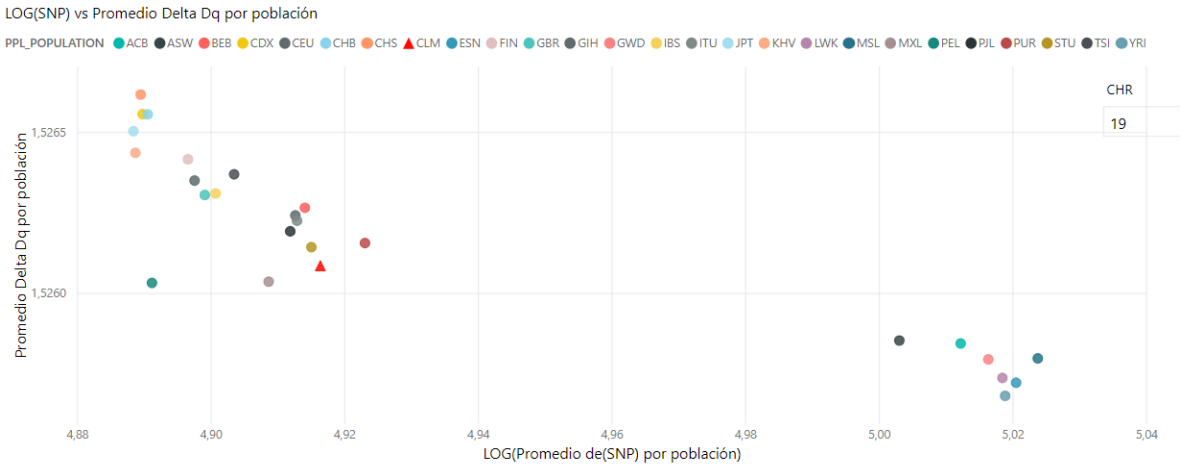


Figura 100: Valores de Δ Dq vs Log(SNP) del cromosoma 20 por población. Colombia representado con un triángulo rojo.

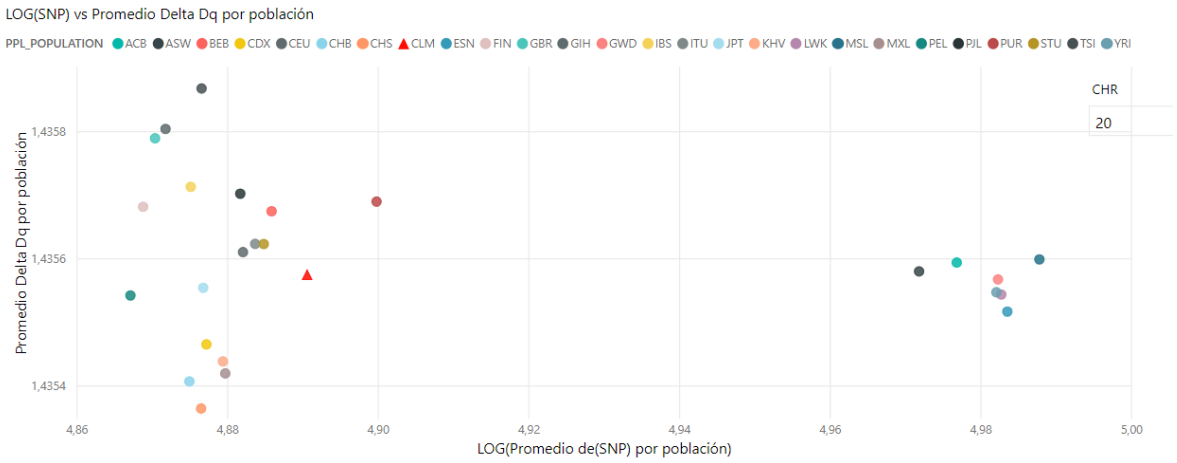


Figura 101: Valores de Δ Dq vs Log(SNP) del cromosoma 21 por población. Colombia representado con un triángulo rojo.

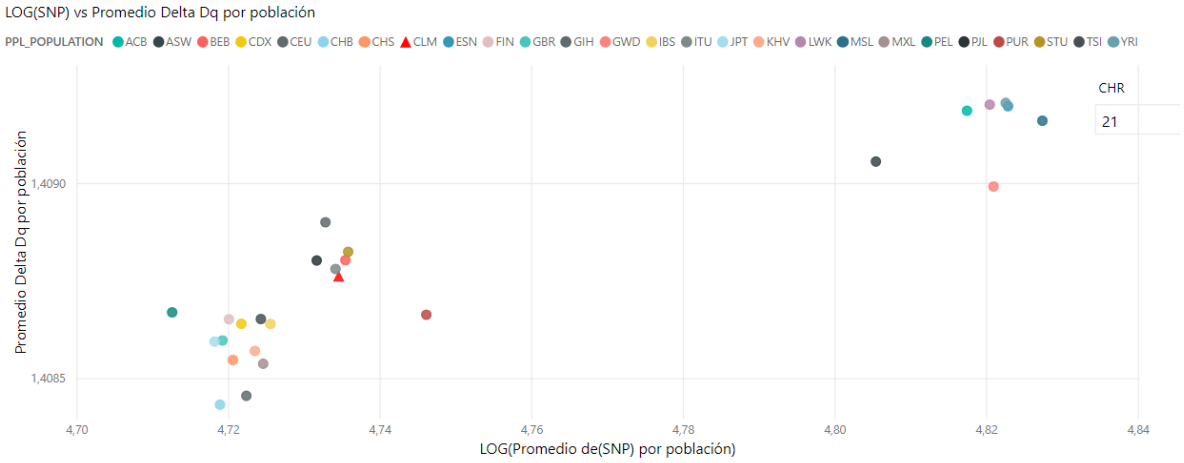


Figura 102: Valores de Δ Dq vs Log(SNP) del cromosoma 22 por población. Colombia representado con un triángulo rojo.

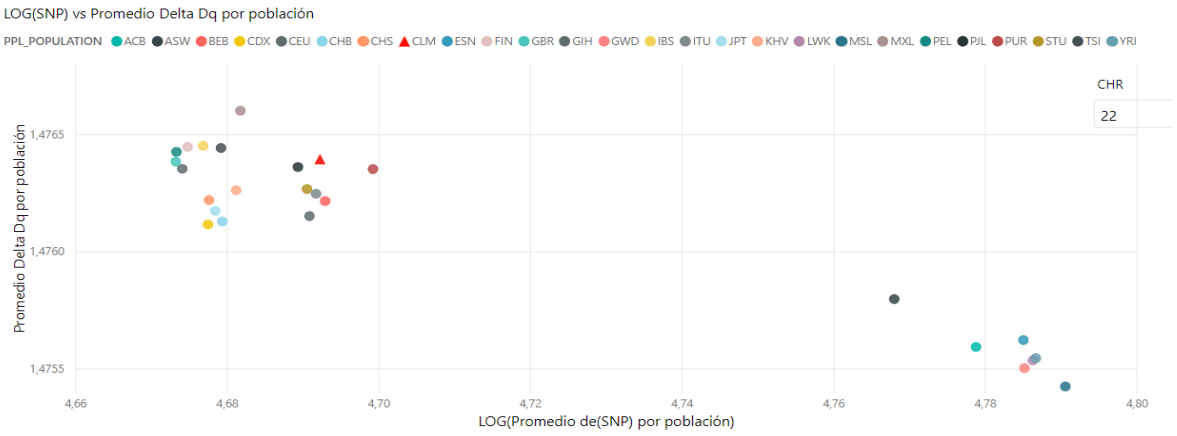
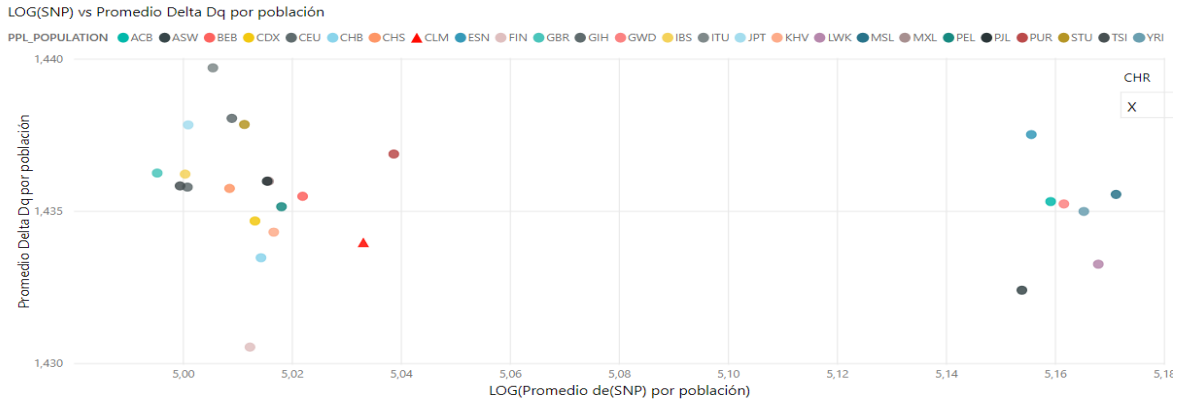
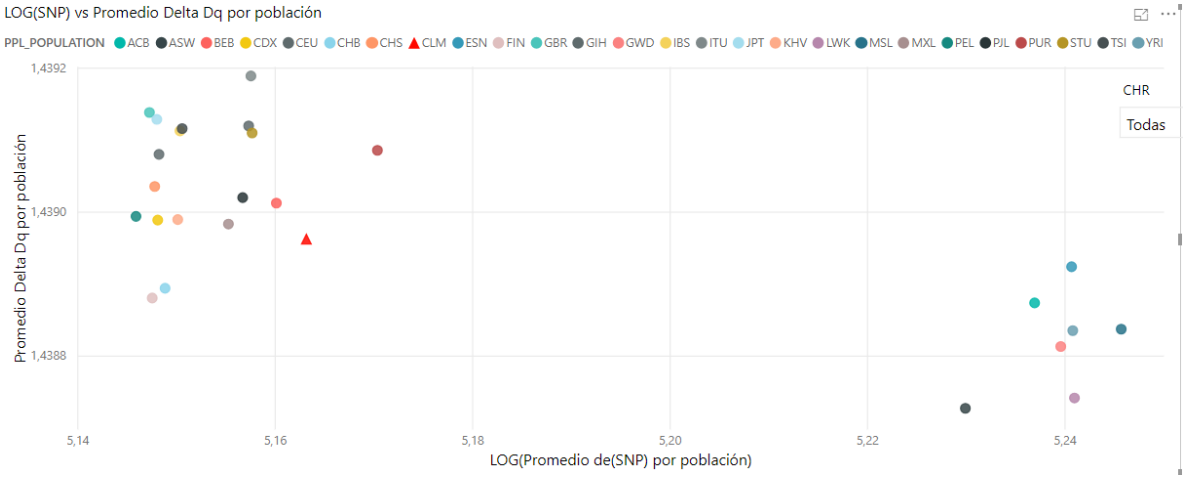


Figura 103: Valores de Δ Dq vs Log(SNP) del cromosoma X por población. Colombia representado con un triángulo rojo.



Para tener un panorama general de las distancias entre las poblaciones, se consolido en un gráfico los promedios ΔDq y de SNP de todos los cromosomas de cada población. Se observa a la población colombiana en la parte baja y derecha de la nube de la izquierda.

Figura 104: Valores de Δ Dq vs Log(SNP) por población. Colombia representado con un triángulo rojo.



Gráficos de relación entre cantidad de SNP y ΔDq por individuo por cromosoma.

Figura 105: Valores de ΔDq vs Log(SNP) del cromosoma 1 para todos los individuos.

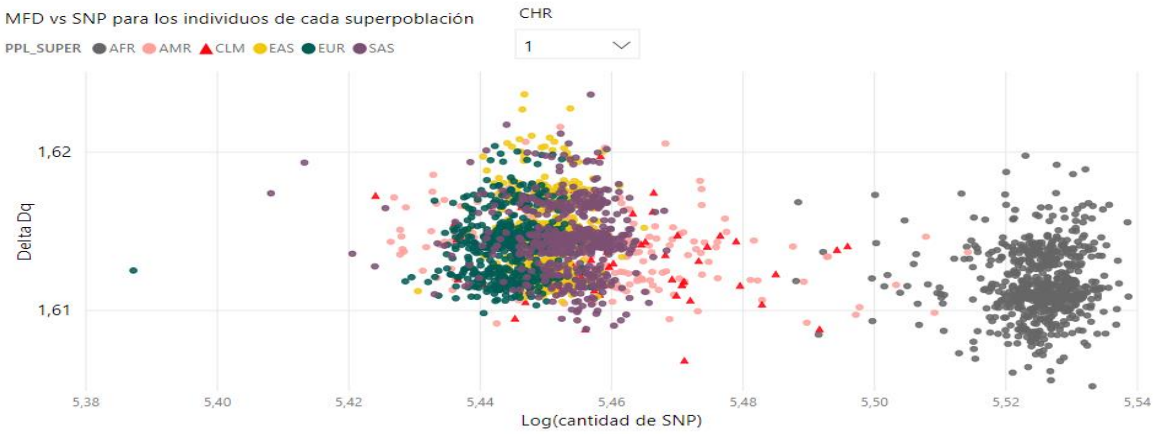


Figura 106: Valores de ΔDq vs Log(SNP) del cromosoma 2 para todos los individuos.

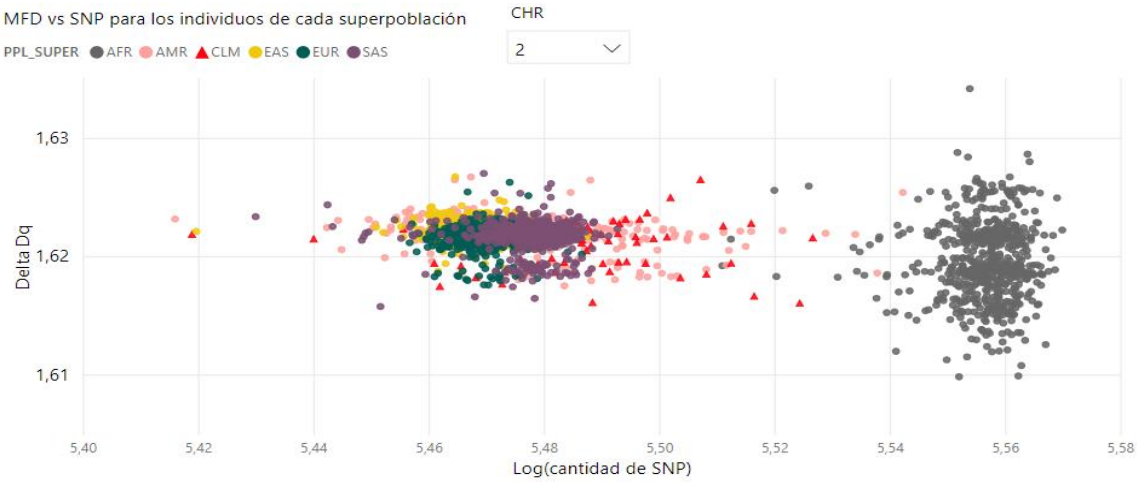


Figura 107. Valores de ΔDq vs Log(SNP) del cromosoma 3 para todos los individuos.

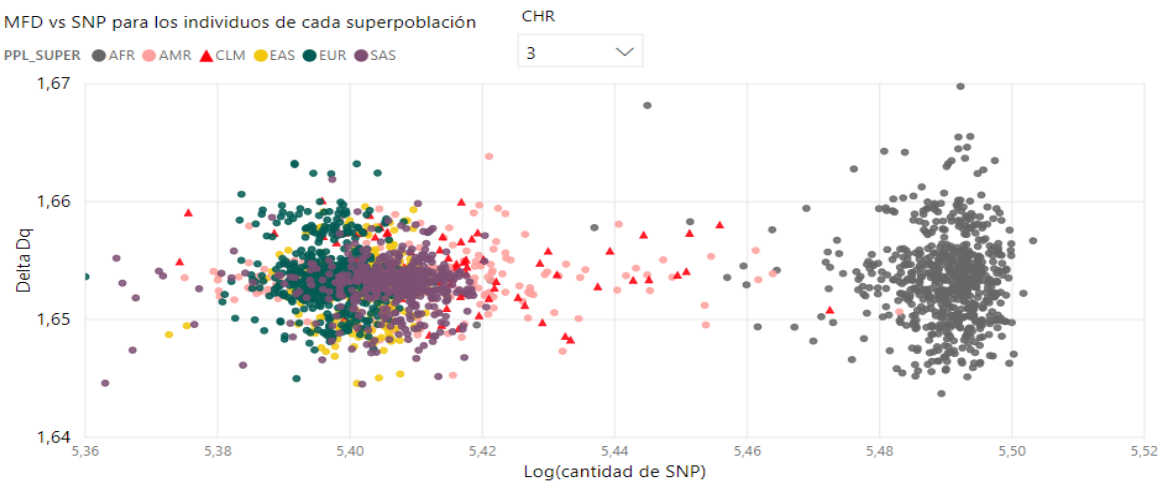


Figura 108: Valores de ΔDq vs Log(SNP) del cromosoma 4 para todos los individuos.

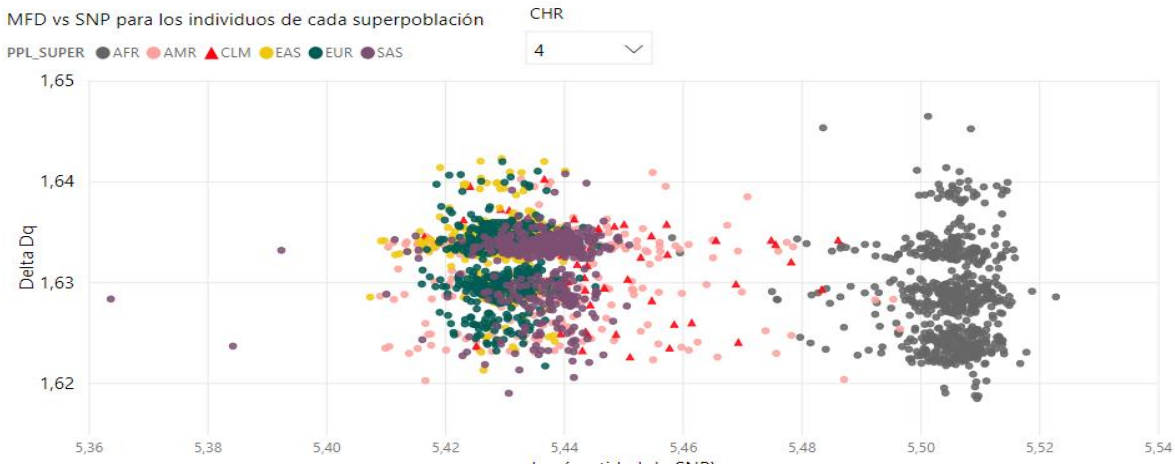


Figura 109: Valores de ΔDq vs Log(SNP) del cromosoma 5 para todos los individuos.

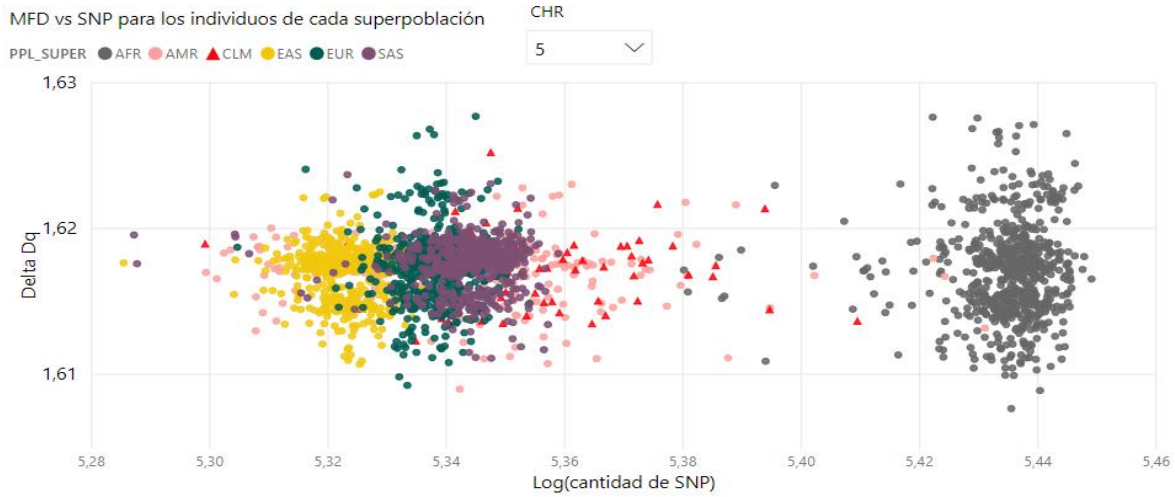


Figura 110. Valores de ΔDq vs Log(SNP) del cromosoma 6 para todos los individuos.

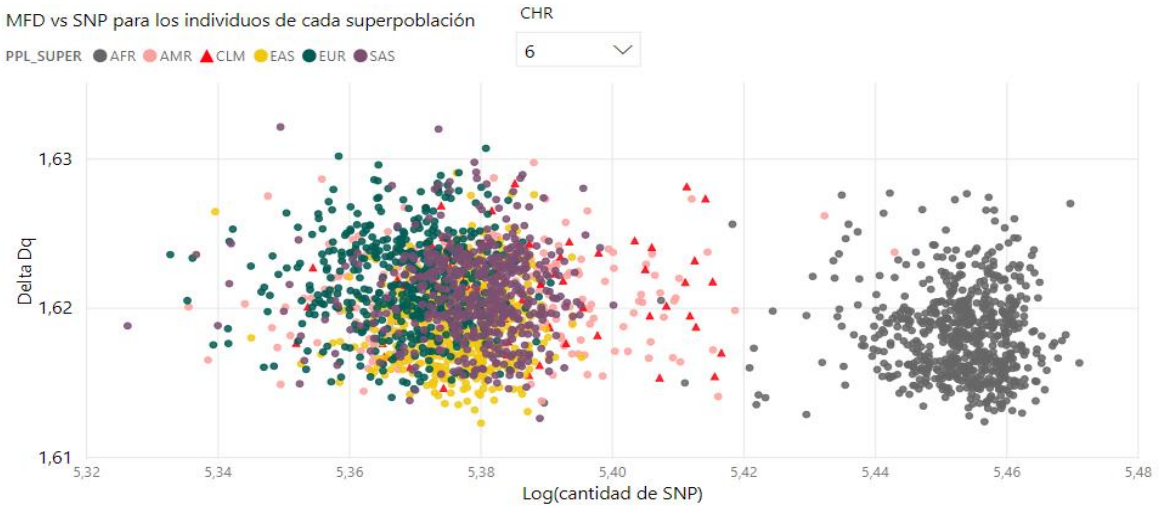


Figura 111: Valores de ΔDq vs Log(SNP) del cromosoma 7 para todos los individuos.

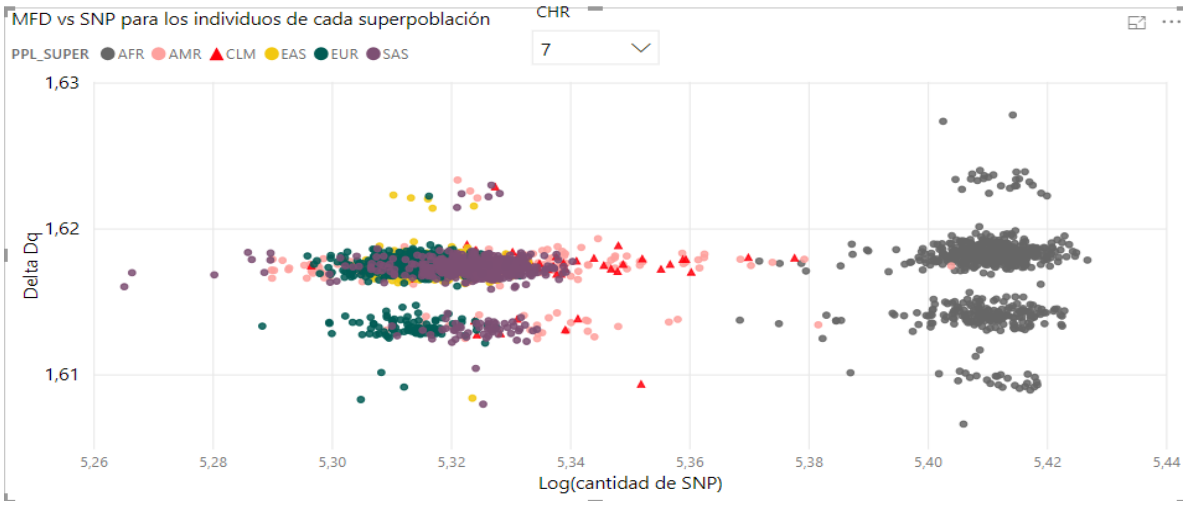


Figura 112: Valores de ΔDq vs Log(SNP) del cromosoma 8 para todos los individuos.

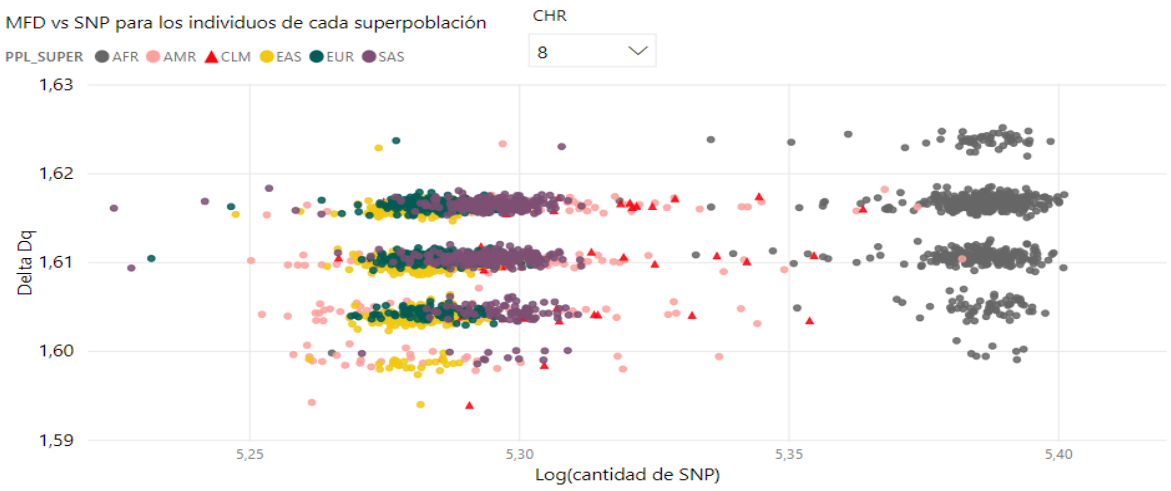


Figura 113: Valores de ΔDq vs Log(SNP) del cromosoma 9 para todos los individuos.

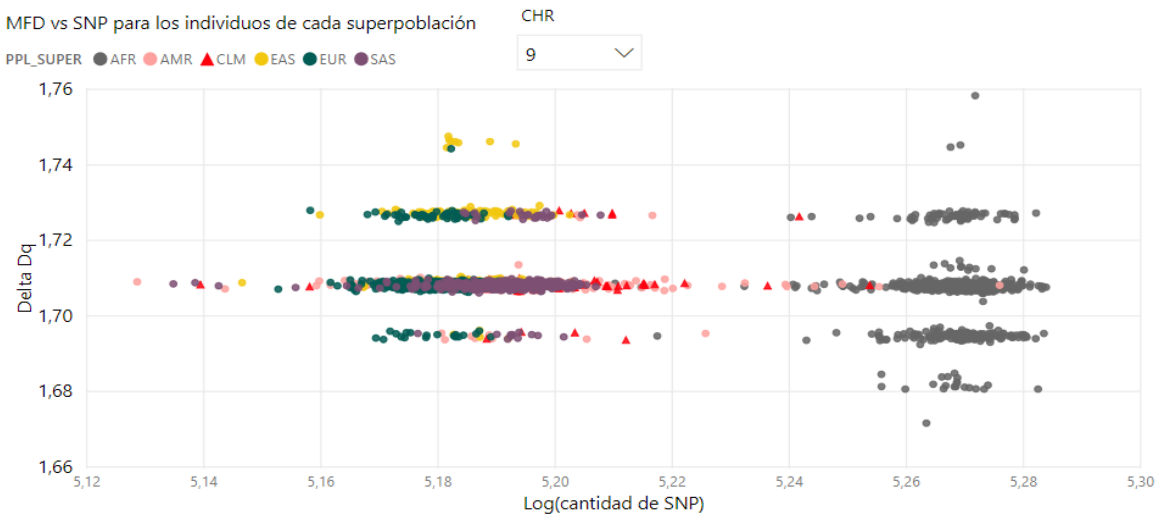


Figura 114: Valores de ΔDq vs Log(SNP) del cromosoma 10 para todos los individuos.

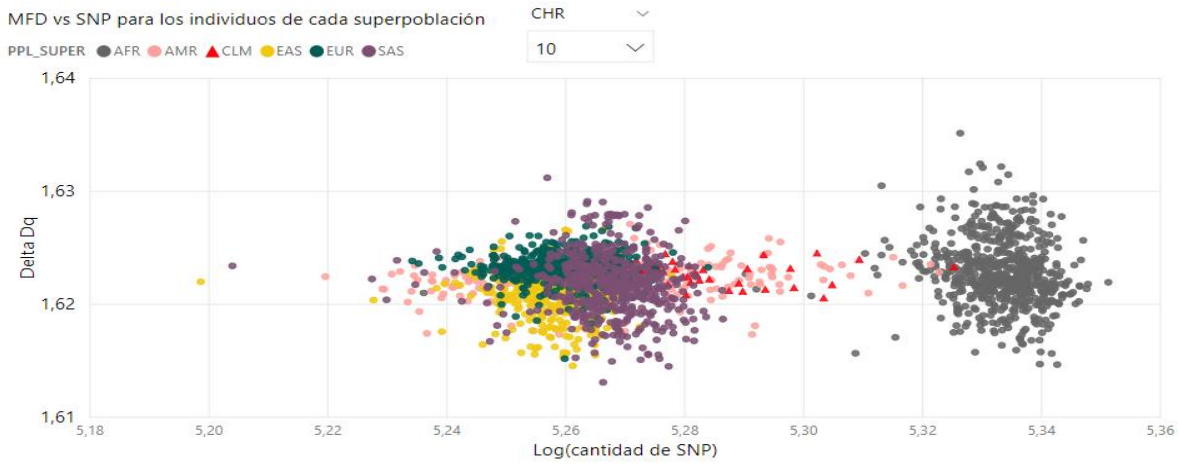


Figura 115: Valores de ΔDq vs Log(SNP) del cromosoma 11 para todos los individuos.

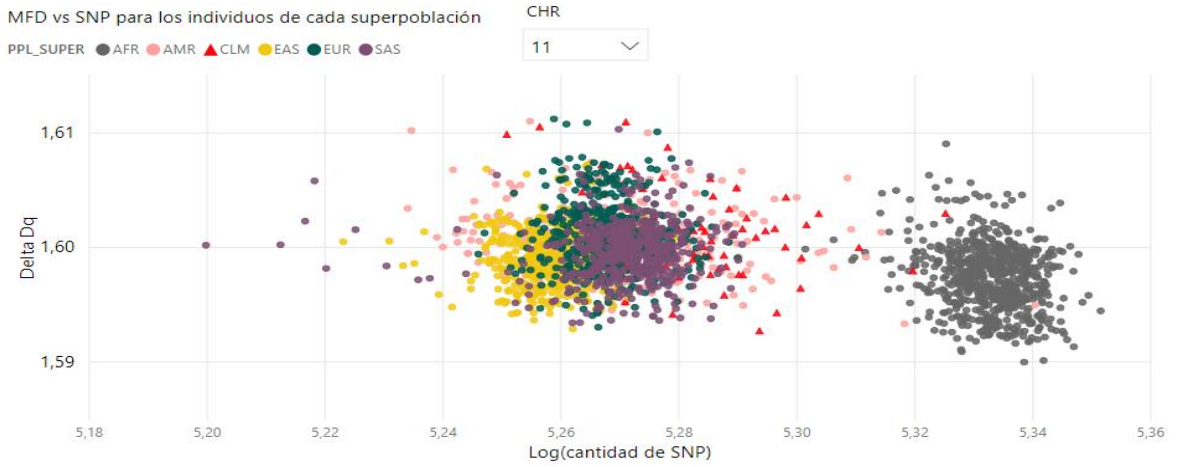


Figura 116: Valores de ΔDq vs Log(SNP) del cromosoma 12 para todos los individuos.

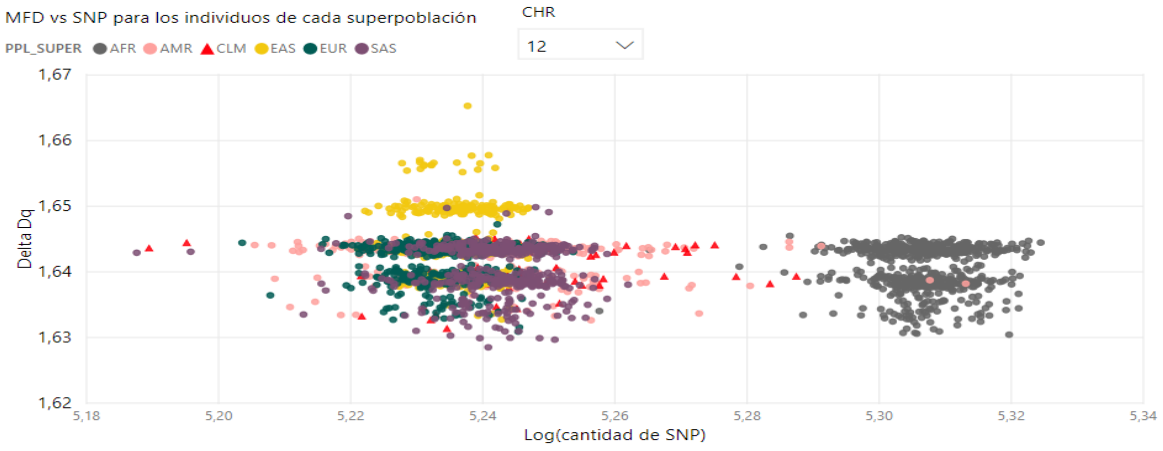


Figura 117: Valores de ΔDq vs Log(SNP) del cromosoma 13 para todos los individuos.

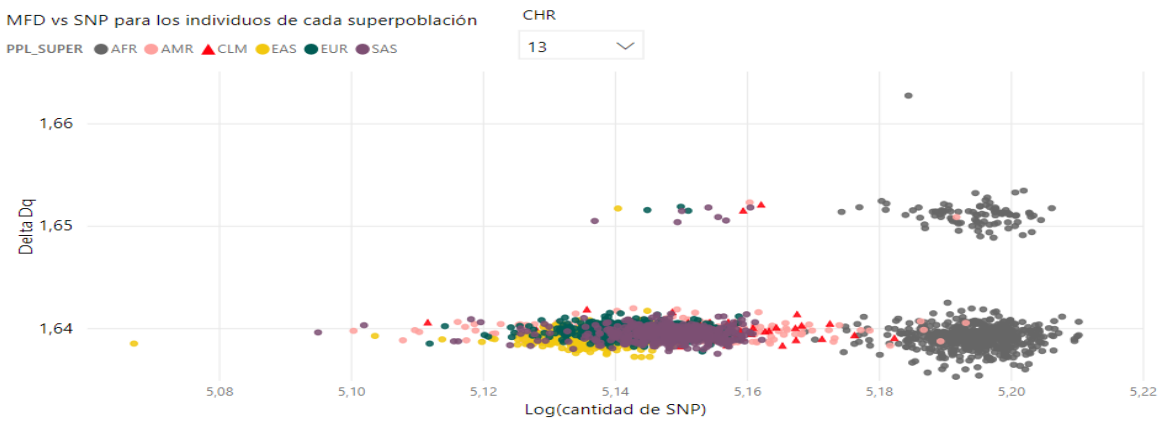


Figura 118: Valores de ΔDq vs Log(SNP) del cromosoma 14 para todos los individuos.

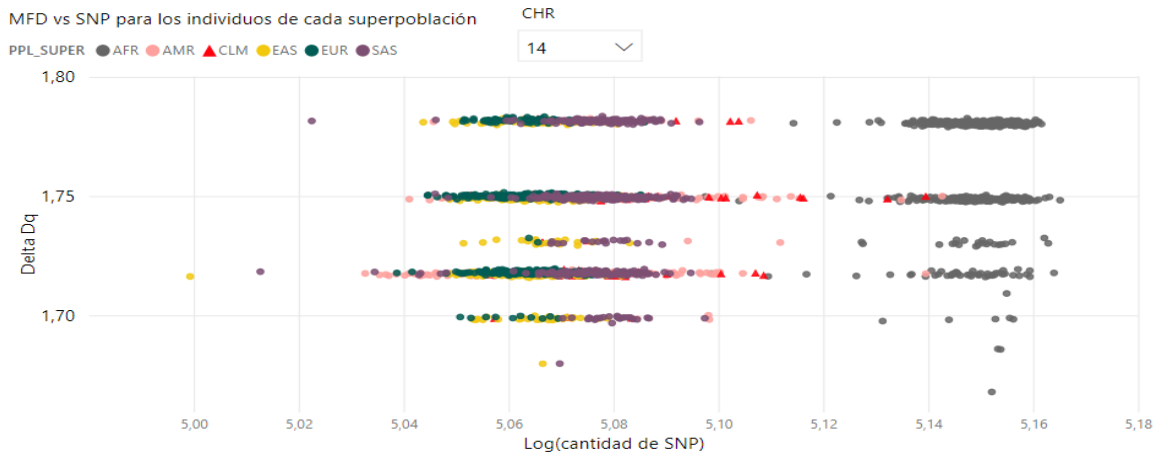


Figura 119: Valores de ΔDq vs Log(SNP) del cromosoma 15 para todos los individuos.

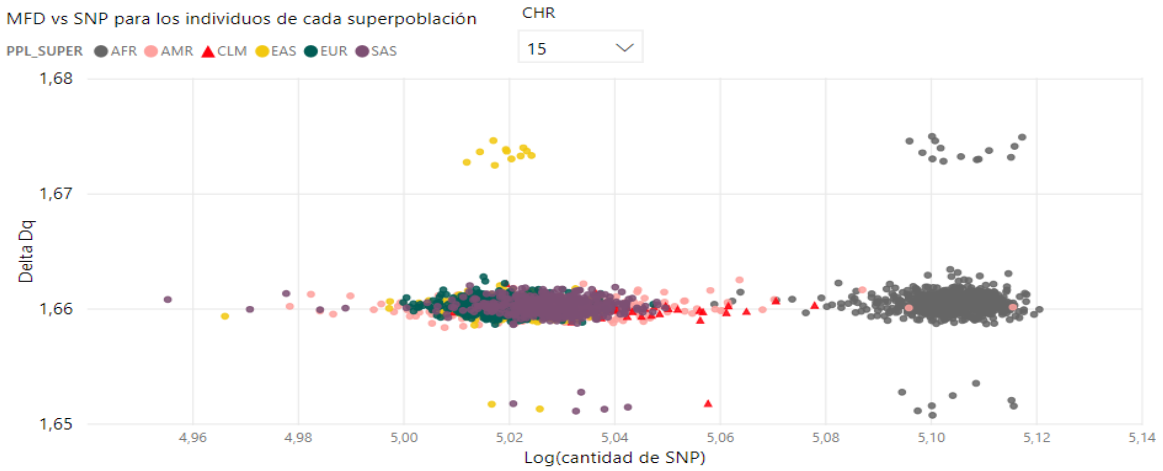


Figura 120: Valores de ΔDq vs Log(SNP) del cromosoma 16 para todos los individuos.

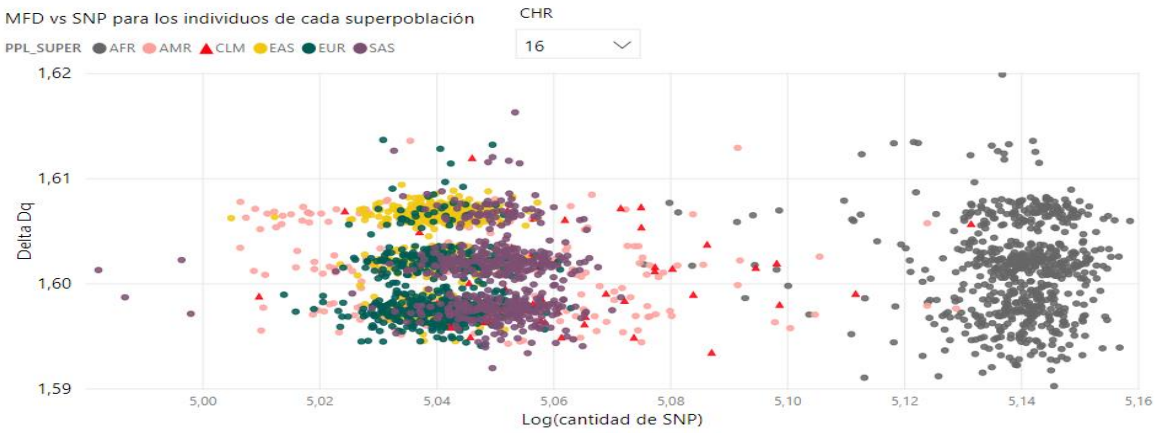


Figura 121: Valores de ΔDq vs Log(SNP) del cromosoma 17 para todos los individuos.

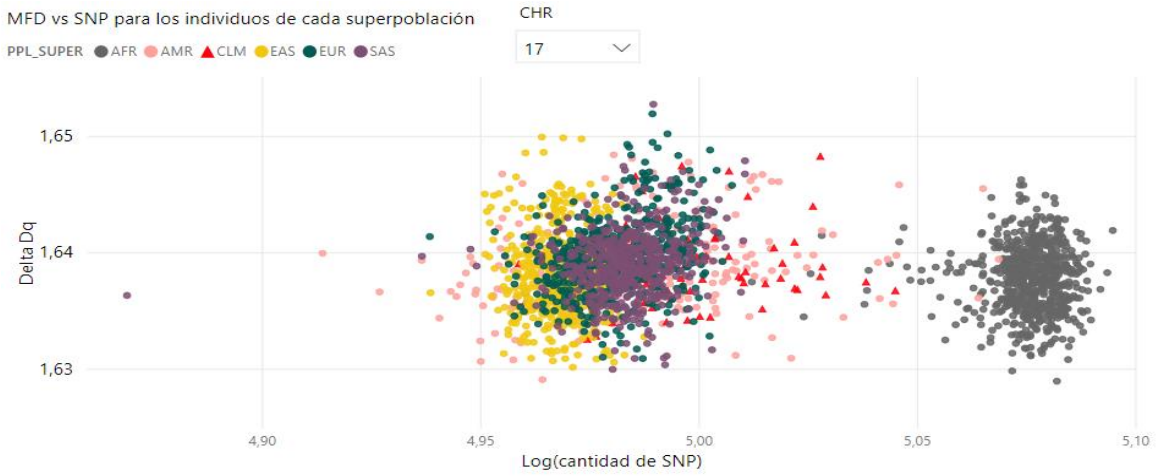


Figura 122: Valores de ΔDq vs Log(SNP) del cromosoma 18 para todos los individuos.

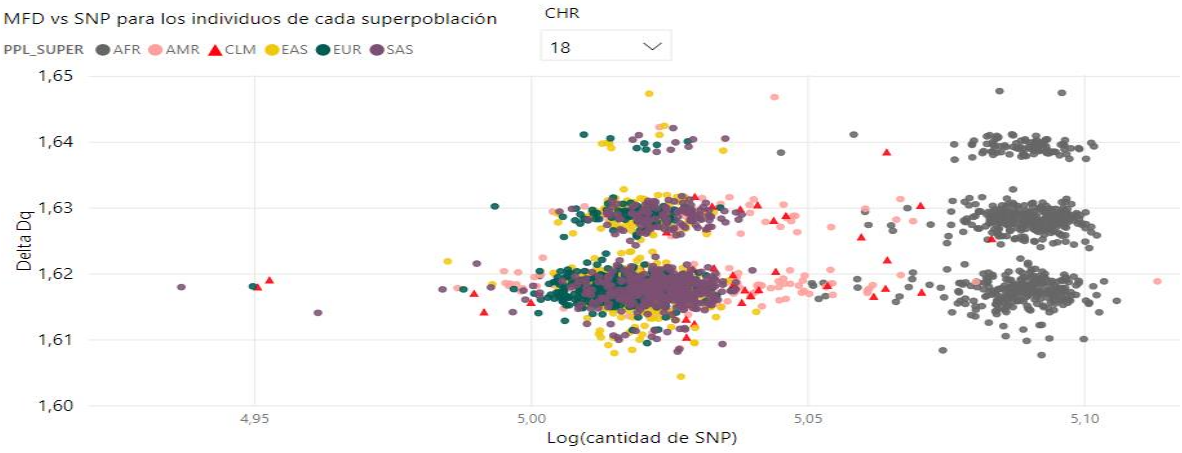


Figura 123: Valores de ΔDq vs Log(SNP) del cromosoma 19 para todos los individuos.

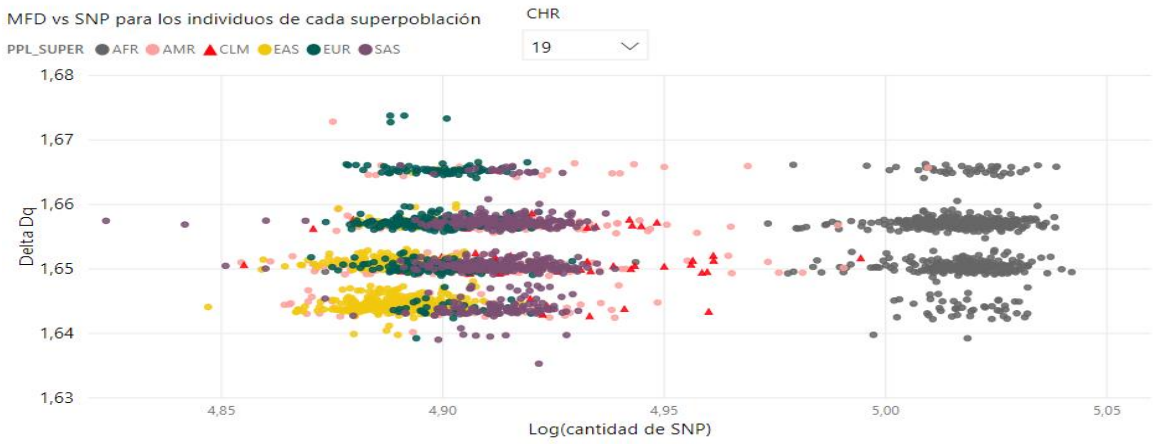


Figura 124: Valores de ΔDq vs Log(SNP) del cromosoma 20 para todos los individuos.

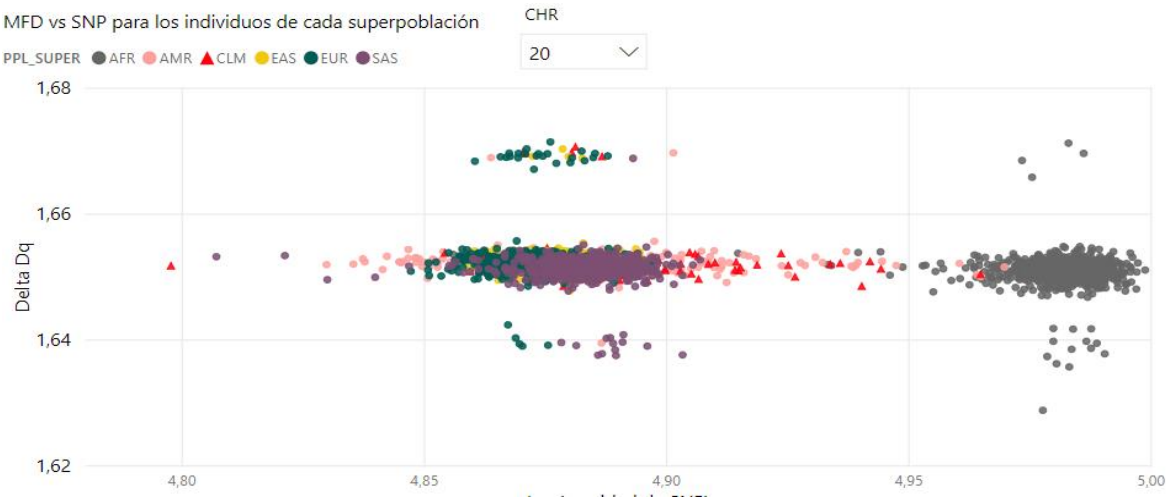


Figura 125: Valores de ΔDq vs Log(SNP) del cromosoma 21 para todos los individuos.

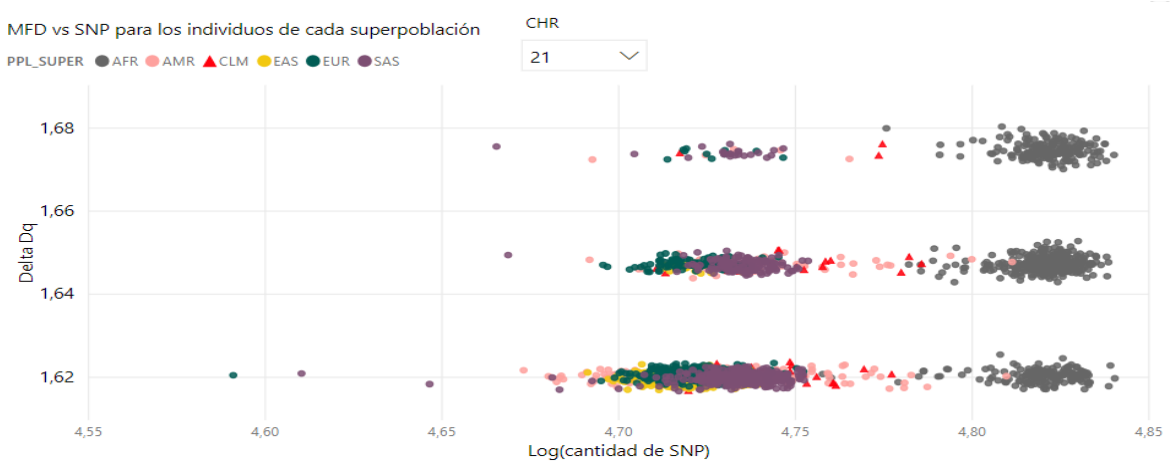


Figura 126: Valores de ΔDq vs Log(SNP) del cromosoma 22 para todos los individuos.

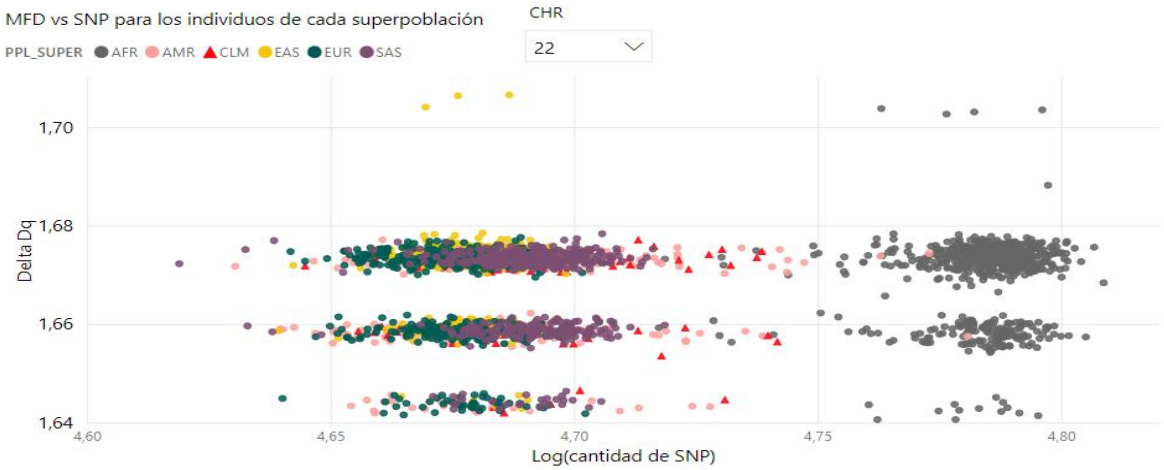
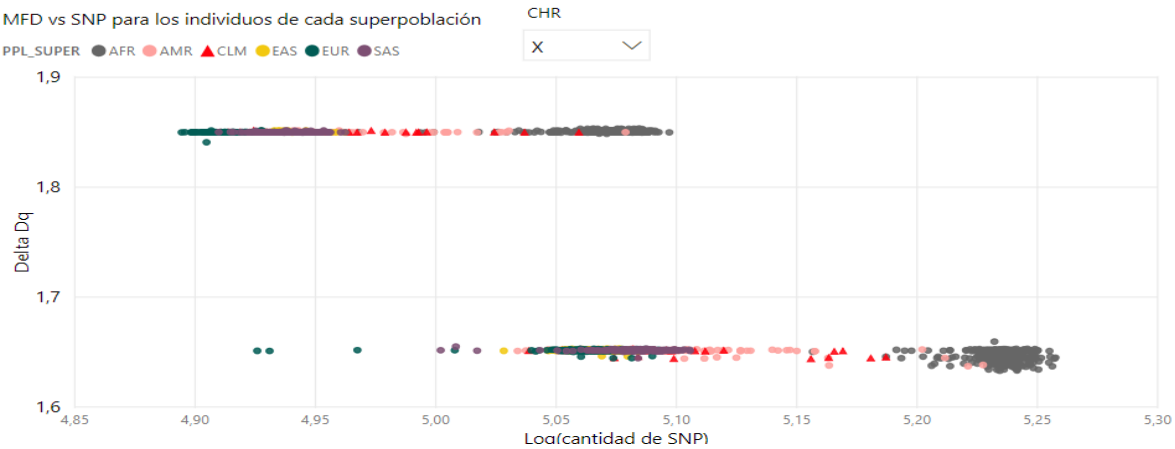


Figura 127: Valores de ΔDq vs Log(SNP) del cromosoma X para todos los individuos.



Revisión de enfermedades con asociadas a las zonas de menor MF en la población colombiana:

Miocardiopatía hipertrófica

Dentro de las enfermedades con baja multifractalidad relativa se encuentra la miocardiopatía hipertrófica, la cual es catalogada como una de las enfermedades cardiacas de origen genético más común cuya prevalencia fue estimada por muchos años de 1:500, pero en años recientes se ha estimado de 1:200[95]. Esta enfermedad se caracteriza por una hipertrofia del ventrículo izquierdo,

habitualmente dicha cavidad es pequeña con una fracción de eyección ventricular izquierda aumentada, y más tarde la rigidez del ventrículo conlleva a una disfunción diastólica que es responsable de la mayoría d ellos signos y síntomas de insuficiencia cardíaca[96].

Se considera que la MCH es una enfermedad del sarcómero ya que todos los genes responsables de la MCH identificados codifican proteínas sarcoméricas. La MCH se hereda de un patrón autosómico dominante. La MCH manifiesta una heterogeneidad genética, con múltiples loci (8 genes), identificándose para cada gen múltiples mutaciones.

Tabla 21: Loci cromosómicos para la miocardiopatía hipertrófica[96]

Gen	*Locus*	Frecuencia	N.º de mutaciones	Tipo de mutaciones
β-MCHC	14q1	35-50%	> 85	Sentido erróneo
MBPC	11q11	15-20%	> 15	Deleciones
Troponina cardíaca T	1q3	15-20%	> 20	Sentido erróneo
Alfa-tropo				
Miosina	15q2	< 5%	5	Sentido erróneo
Troponina cardíaca I	19q13	< 1%	3	
MLC-1	3p	< 1%	2	
MLC-2	12q	< 1%	2	
Actina cardíaca	15q11	(?)	2	
Desconocida	7q3	(?)	(?)	
Desconocida	(?)	(?)	(?)	(?)

De los genes anteriores, el detectado en las ventanas de baja multifractalidad para Colombia fue el MYL2 (Miosina), el cual solo presenta una frecuencia del 5% como causa de la enfermedad.

Diabetes mellitus tipo 1

La diabetes mellitus tipo 1 es una enfermedad crónica en la cual las células del páncreas encargadas de fabricar insulina se destruyen y dejan de generarla, suele tener una aparición brusca. Los estudios de asociación genómica (GWAS) han identificado más de 60 regiones en el genoma humano con relación con la enfermedad. El mayor riesgo heredable esta atribuido al leucocito antígeno

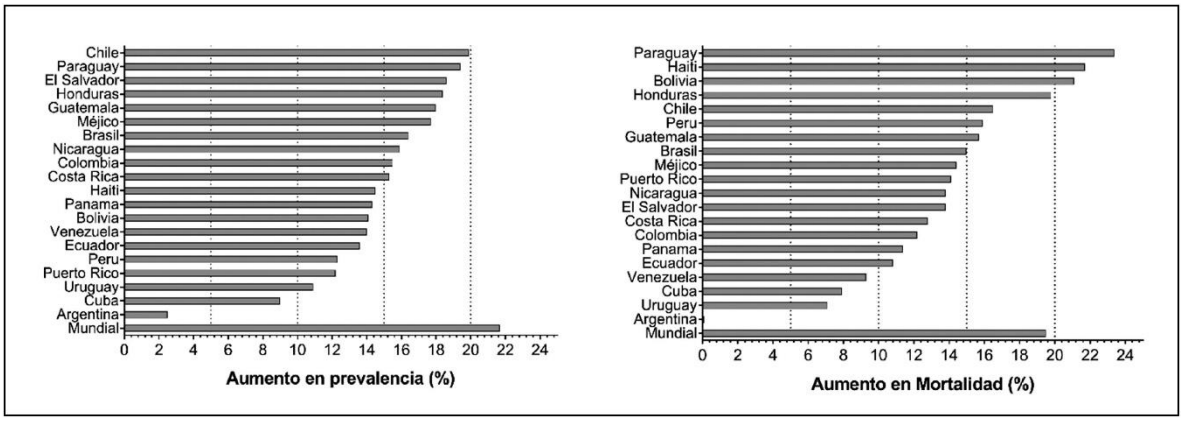
humano (HLA), sin embargo, muchas de las variantes genéticas identificadas caen fuera de las regiones codificantes, lo que hace más difícil su análisis.

En el índice mundial reportado por indexmuldi[97], los Colombia aparece en el puesto 86 de 194 países, con un índice de 7.44, esto la ubica justo en un punto medio frente a los demás países del mundo.

Revisión de enfermedades con asociadas a las zonas de mayor MF en la población colombiana.

Según la teoría propuesta, significaría que en Colombia debería haber una menor prevalencia en las enfermedades cuyos genes estén localizados en regiones de altas MF. Por ejemplo, en el caso de la enfermedad de Parkinson la prevalencia a nivel de Latinoamérica para Colombia es intermedia, si comparamos ésta, con otros países a nivel mundial o con Chile quienes muestran las más altas prevalencias, Figura 27. Igualmente, es intermedia si comparamos éstas con los individuos tomados en el presente estudio, como europeos, asiáticos, africanos y los individuos de México, Perú y Puerto Rico.

Figura 128: Rankin de prevalencia de la enfermedad de Parkinson en Latinoamérica. El mundo y Chile lideran el ranking. Rev. méd. Chile vol.147 no.4 Santiago abr. 2019.



Melanoma

La capa más profunda de la epidermis, ubicada justo encima de la dermis, contiene células llamadas melanocitos. Los melanocitos producen el pigmento o el color de la piel. El melanoma comienza cuando los melanocitos sanos cambian y crecen sin control, formando un tumor canceroso. Un tumor canceroso es maligno,

lo que significa que puede crecer y diseminarse a otras partes del cuerpo. A veces, el melanoma se desarrolla a partir de un lunar normal que una persona ya tiene en su piel. Cuando esto sucede, el lunar sufrirá cambios que generalmente se pueden ver, como cambios en la forma, el tamaño, el color o el borde del lunar.

El melanoma puede desarrollarse en cualquier parte del cuerpo, incluida la cabeza y el cuello, la piel debajo de las uñas, los genitales e incluso las plantas de los pies o las palmas de las manos.[98]

La susceptibilidad de ciertos individuos se da por mutaciones principalmente en los genes CDKN2A, CDK4, BAP1, POT1, ACD, TERF2IP y TERT, a pesar de eso, solo el 50% de los casos pueden ser vinculados a causas genéticas. En la ventana que se detectó la alta multifractalidad contiene el gen ARID1B el cual participa en el proceso de remodelamiento de la cromatina[99].

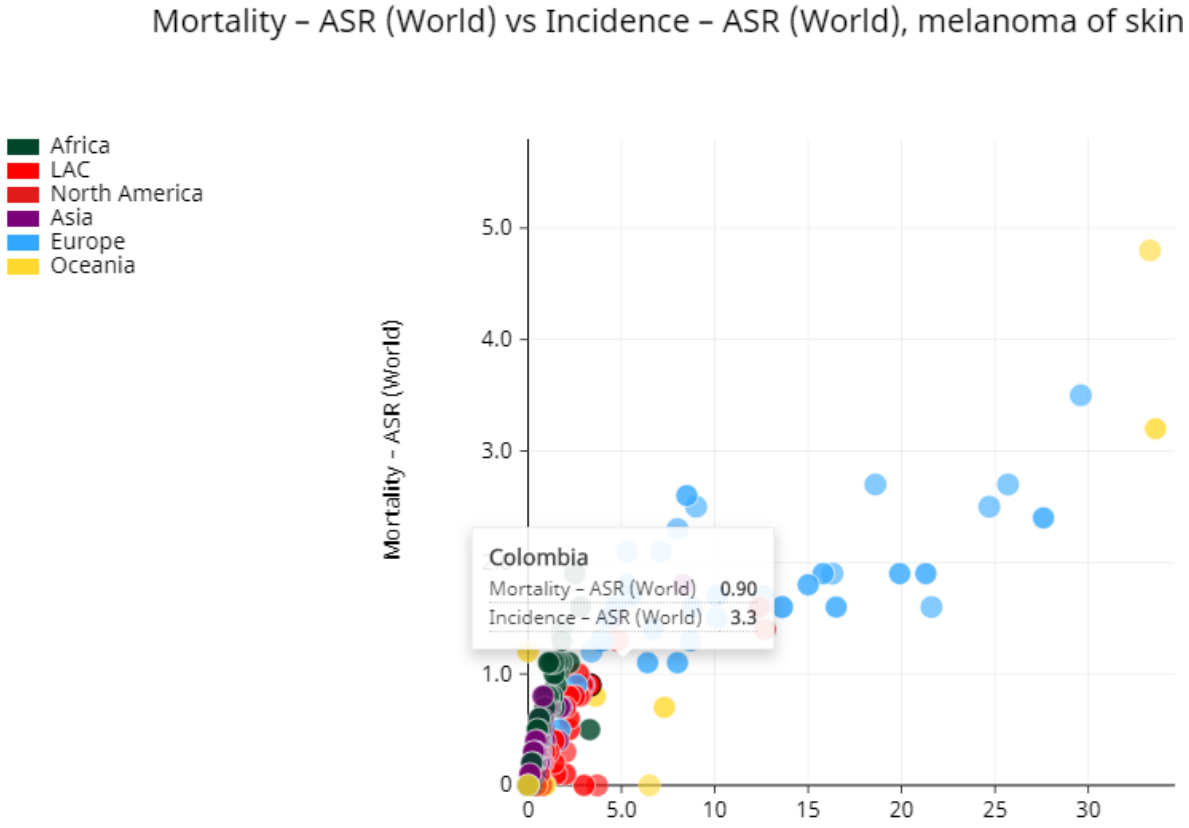
La incidencia de melanoma en el mundo la lidera Australia seguido por Nueva Zelanda, el primer país latinoamericano es Uruguay en el puesto 40. Colombia ocupa la casilla 50 entre 185 países (Tabla 22).

Tabla 22: Incidencia de melanoma en el mundo[100]

Puesto	Población	Incidencia	Mortalidad
1	Australia	33,6	3,2
2	New Zealand	33,3	4,8
3	Norway	29,6	3,5
4	Denmark	27,6	2,4
5	The Netherlands	25,7	2,7
6	Sweden	24,7	2,5
7	Germany	21,6	1,6
8	Switzerland	21,3	1,9
9	Belgium	19,9	1,9
	…	…	…
40	Uruguay	4,8	1,3
41	Russian Federation	4,7	1,6
42	Cyprus	4,5	1,5
43	Republic of	4,3	1,5

	Moldova		
44	Bulgaria	4	1,3
	…	…	…
50	Colombia	3,3	0,9

Figura 129: Prevalencia vs mortalidad de melanoma[100]



Cáncer gástrico

El cáncer de estómago, también llamado cáncer gástrico, comienza cuando las células sanas del estómago se vuelven anormales y crecen sin control. El cáncer puede comenzar en cualquier parte del estómago y se puede propagar a los ganglios linfáticos cercanos y a otras partes del cuerpo, como el hígado, los huesos, los pulmones y los ovarios de una mujer. La mayoría de los cánceres de estómago son un tipo llamado adenocarcinoma. Esto significa que el cáncer comenzó en el tejido glandular que recubre el interior del estómago. Otros tipos de tumores cancerosos que se forman en el estómago incluyen linfoma, sarcoma gástrico y tumores neuroendocrinos, pero estos son poco frecuentes. Los factores

de riesgo incluyen fumar, infección con la bacteria H. pylori y ciertas afecciones hereditarias.

Existen cuatro grupos de genes involucrados en el desarrollo de cáncer. Los genes supresores de tumor (GST), los protooncogenes, los genes de reparación del ADN (son GST, pero su única función se asocia con la reparación y mantenimiento del genoma) y los genes proapoptóticos - son protooncogenes, solo que su función está asociada con la muerte celular programada.

El gen relacionado con la ventana de alta multifractalidad es el CDX2, el cual ha sido vinculado a la inhibición del crecimiento de las células del cáncer gástrico[101].

El cáncer gástrico (CG) ocupa el quinto puesto en incidencia y el tercero en mortalidad en el mundo. Es la primera causa de muerte por cáncer en Colombia. En la actualidad, el CG es un problema de salud pública en el país por tres factores: las altas tasas de incidencia/mortalidad, los sobrecostos de los tratamientos de la enfermedad en estadios avanzado y la falta de acceso de la población a los servicios de salud.

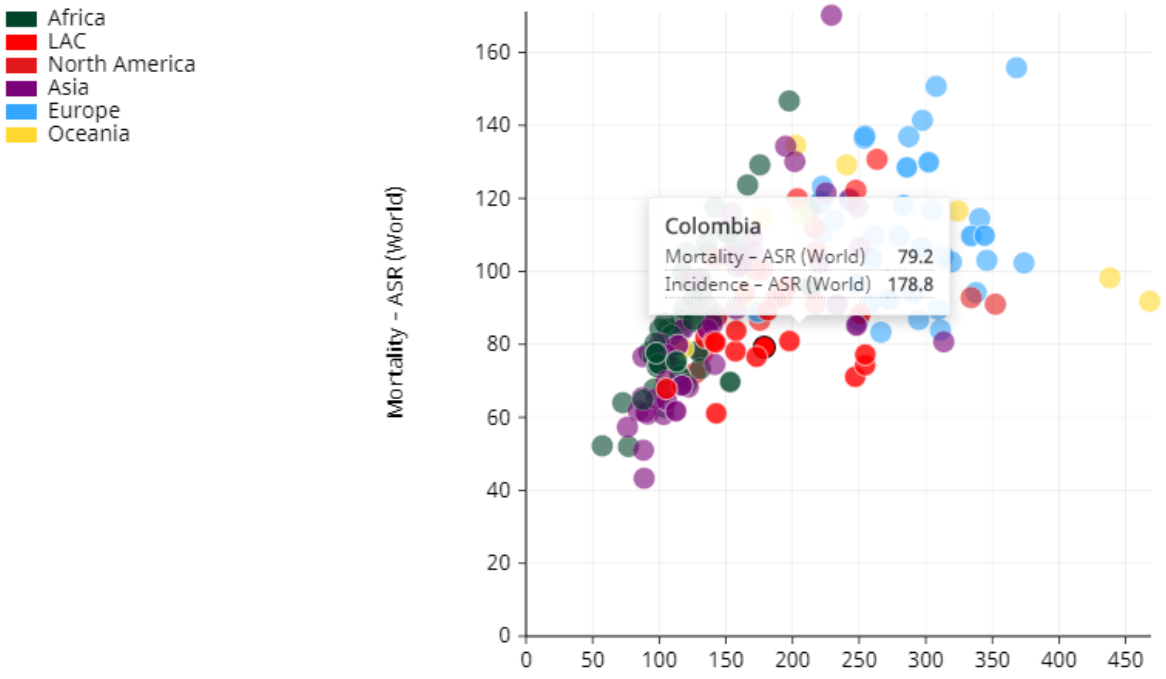
El país con mayor prevalencia de cáncer gástrico es la república de Corea seguido por Mongolia, Japón y China, el primer país latinoamericano es Chile en el puesto 7 seguido de cerca por Perú en el puesto 9. Colombia se encuentra en el puesto 26, mucho más alto que los otros tipos de cáncer donde Colombia tiene una alta multifractalidad (Tabla 23).

Tabla 23: Incidencia y mortalidad de cáncer gástrico[102]

Puesto	Población	Incidencia	Mortalidad
1	Korea, Republic of	39,6	7
2	Mongolia	33,1	25
3	Japan	27,5	9,5
4	China	20,7	17,5
5	Bhutan	19,4	18,9
6	Kyrgyzstan	18,6	16,6
7	Chile	17,8	11,5
8	Belarus	16,5	9,6
9	Peru	16,1	12,8
…	…	…	…
16	Guatemala	14,4	12,1

17	Ecuador	13,8	10,8
18	Brunei	13,6	8,8
19	Costa Rica	13,4	10,1
20	Lithuania	13,3	10,6
21	Russian Federation	13,3	10,8
22	Republic of Moldova	13	10
23	French Guyana	13	6
24	Latvia	12,9	9,5
25	Myanmar	12,8	10,7
26	Colombia	12,8	9,4

Figura 130: Incidencia y mortalidad de cáncer gástrico[102]



Leucemia

La leucemia es un cáncer de la sangre. Comienza cuando las células sanguíneas sanas cambian y crecen sin control. La leucemia mieloide aguda (AML) es un tipo de leucemia que es un cáncer del tejido formador de sangre en la médula ósea, el tejido esponjoso dentro de los huesos. La médula ósea produce muchas células

cancerosas anormales, también llamadas blastos o mieloblastos porque se parecen a las células blásticas inmaduras sanas. En lugar de convertirse en células sanguíneas maduras sanas, las células cancerosas se dividen rápidamente y fuera de control. Finalmente, estos mieloblastos llenan la médula ósea, evitando que se formen células sanas y luego se acumulan en el torrente sanguíneo. También pueden moverse hacia los ganglios linfáticos, el cerebro, la piel, el hígado, los riñones, los ovarios (en las niñas), los testículos (en los niños) y otros órganos[103]. El gen FLT3 el cual es vinculado a la ventana identificada como de alta multifractalidad, normalmente se utiliza como marcador para la presencia de la enfermedad cuando este se encuentra conmutaciones, normalmente esta condición no se hereda, sino que se desarrolla debido a múltiples factores de riesgo.

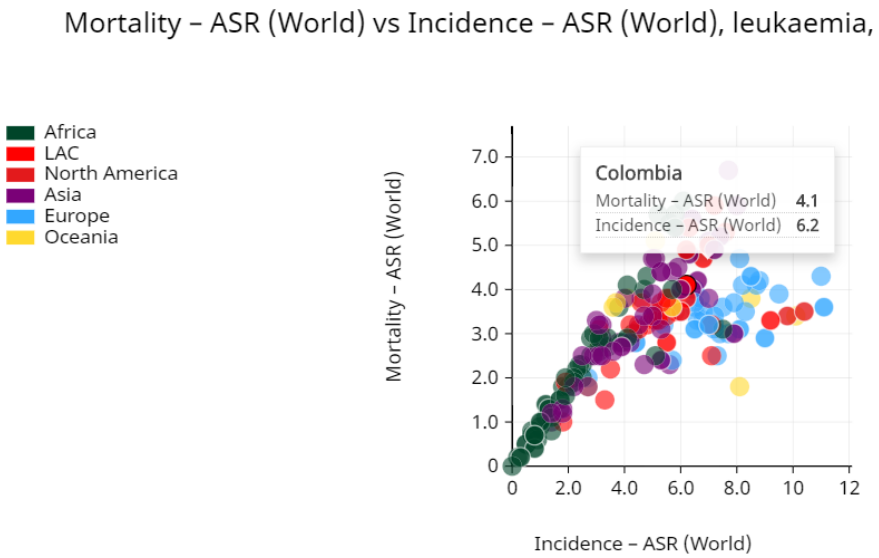
En la prevalencia de leucemia, la clasificación la lidera Bélgica, seguida por Lituania y Canadá, el primer país latinoamericano en el ranking es Perú en el puesto 23 y con una incidencia de 7,6. Colombia ocupa el puesto 56 entre 185 países, superado en Suramérica solo por Perú y Ecuador (Tabla 24).

Tabla 24: Incidencia de Leucemia en el mundo en 2018[104]

Puesto	Población	Incidencia	Mortalidad
1	Belgium	11,1	3,6
2	Lithuania	11	4,3
3	Canada	10,4	3,5
4	Australia	10,1	3,4
5	United States of America	9,8	3,4
6	France	9,5	3,9
7	France, Martinique	9,2	3,3
8	United Kingdom	9	2,9
9	Latvia	8,8	4,2
	…	…	…
23	Peru	7,6	5,3
24	Denmark	7,5	3,6
25	Francia	7,5	3,1
26	Jordan	7,4	5,2
27	Switzerland	7,4	3
28	Finland	7,3	2,5
29	El Salvador	7,2	5,9

	…		…
56	Colombia	6,2	4,1

Figura 131: Incidencia y mortalidad por leucemia en el mundo[104]



Cáncer de tiroides

El cáncer de tiroides comienza cuando las células sanas en la tiroides cambian y crecen sin control, formando una masa llamada tumor. La glándula tiroides contiene 2 tipos de células. Las Células foliculares responsables de la producción de hormona tiroidea. La hormona tiroidea es necesaria para vivir pues controla el metabolismo básico del cuerpo, la rapidez con que se queman las calorías. Esto puede afectar la pérdida de peso y el aumento de peso, ralentizar o acelerar los latidos del corazón, aumentar o disminuir la temperatura corporal, influir en la rapidez con que los alimentos se mueven a través del tracto digestivo, controlar la forma en que los músculos se contraen y controlar la rapidez con que se reemplazan las células moribundas. Y el segundo tipo son las células C, que producen calcitonina, una hormona que participa en el metabolismo del calcio. Otros tipos de cáncer pueden comenzar dentro o alrededor de la glándula tiroides.

El gen BRAF produce una proteína que ayuda a controlar la reproducción celular. Se le conoce como un oncogen, el cual activa la reproducción de las células según

sea necesario. Cuando se encuentra mutado no puede evitar que las células se reproduzcan y dicha reproducción descontrolada puede causar cáncer. Por esa razón también está vinculado con otros tipos de cáncer.

La incidencia de cáncer de tiroides a nivel mundial la lidera Australia, seguido por Nueva Zelanda e Irlanda, el primer país latinoamericano es Uruguay y Colombia se ubica en el puesto 76 (Tabla 25).

Tabla 25: Incidencia de cáncer de tiroides en el mundo[105]

Puesto	Población	Incidencia	Mortalidad
1	Australia	468	91,8
2	New Zealand	438,1	98,2
3	Ireland	373,7	102,3
4	Hungary	368,1	155,8
5	United States of America	352,2	91
6	Belgium	345,8	103
7	France	344,1	109,8
8	Denmark	340,4	114,5
9	Norway	337,8	94,2
…	…	…	…
31	Uruguay	263,4	130,7
32	Belarus	260,7	109,7
33	Portugal	259,5	103,6
34	Iceland	257,8	91,1
35	France, Guadeloupe	254,6	77,2
36	Puerto Rico	254,5	74,3
…	…	…	…
76	Colombia	178,8	79,2

Figura 132: Incidencia y mortalidad de cáncer de tiroides en el mundo[105]

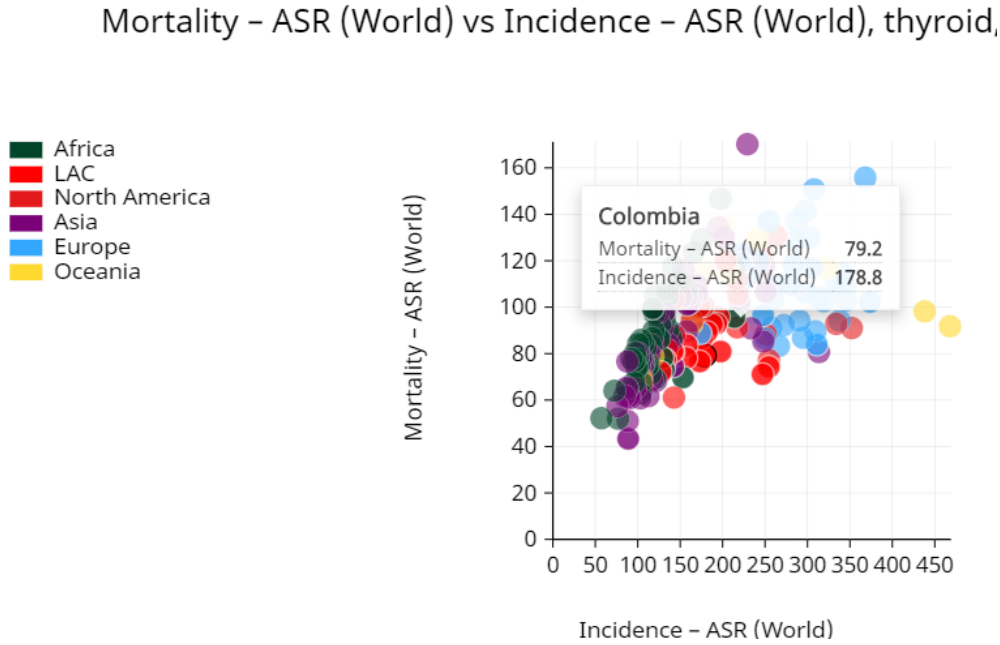


Tabla 26: Relación MFD y Prevalencia de algunas enfermedades

Cod_enfermedad	Enfermedad	MFD	Prev X 100000
H00057	Parkinson disease	1,47086751	4,8
H00031	Breast cancer	1,47482013	44,1
H00408	Type 1 diabetes mellitus	1,46488824	70
H01593	Osteoporosis	1,44822035	69
H00045	Pancreatic neuroendocrine tumor	1,45640762	4
H01457	Diabetic retinopathy	1,48967891	19,76
H00003	Acute myeloid leukemia	1,47537332	6,2
H00032	Thyroid cancer	1,48015758	9
H00038	Melanoma	1,45967463	3,3
H00020	Colorectal cancer	1,46939541	178
H01481	Myelodysplastic syndrome	1,48985035	5
H00247	Multiple endocrine neoplasia syndrome; Wermer syndrome; Sipple syndrome	1,50344472	3
H00018	Gastric cancer	1,48115803	12,8