UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
INSTITUTO DE INGENIERÍA

Titulo:

"Herramienta bioinformática para la estimación de parámetros genéticos de poblaciones, utilizando marcadores SNP."

TESIS PARA OBTENER EL GRADO DE:
MAESTRA EN INGENIERÍA

PRESENTA:
KATHLEEN STEPHANY SOTO SIGALA

DIRECTOR DE TESIS
DR. RAFAEL VILLA ANGULO

Enero, 2019.

Resumen

En Proyecto de tesis de maestría se desarrolló de una herramienta bioinformática, con un ambiente gráfico amigable, que implementa algoritmos para realizar estimación de Frecuencias alélicas, Heterosigocidad y Consanguinidad, partiendo de información de marcadores tipo SNP, de una población de individuos. En el este documento se presentan las etapas desarrolladas para el cumplimiento del objetivo general. Primero se presenta un estudio previo sobre conceptos básico de genética y genética de poblaciones, posteriormente se presenta el diseño de la herramienta bioinformática, incluyendo el código en el lenguaje de programación Python de los algoritmos para estimar frecuencias alélicas, heterosigosidad y consanguinidad. Igualmente se presenta un ejemplo de la interfaz gráfica desarrollada, y resultados obtenidos al probar la herramienta con datos de genotipos reales.

En conclusión, este trabajo presenta los primeros pasos para la generación de una herramienta bioinformática que englobará estimaciones de parámetros genéticos basada en marcadores SNP. En especial se consideraron la estimación de parámetros de Frecuencias alélicas, Heterosigocidad y consanguinidad.

Como trabajo futuro se plantea obtener una representacion gráfica de los datos, de una forma mas amigable para el usuario, y que integres el analisis de otros parametros genéticos, tales como Desequilibrio por ligamento, pruebas de asocacion, estmiacion de Valores Genomicos, entre otros.

Índice

CAPITULO 1. Introducción

1.1. Bioinformática.

Existen diversas formas de definir la Bioinformática debido al amplio rango de aplicaciones de esta disciplina. La Bioinformática involucra las ciencias computacionales y ciencias de la biología. Una definición muy aceptada fue propuesta por Luscombe et al; quien definió a la Bioinformática como la unión de la biología, la informática y las ciencias computacionales, para el almacenamiento, manipulación y análisis de información biológica, y resolver problemas complejos sobre la evolución, la salud y el clico de vida de los seres vivos, desde una perspectiva genómica.

El comienzo de la bioinformática se remota a Margaret Dayhoff, en 1968, y su colección de secuencias de proteínas conocida como Atlas de secuencia y estructura de proteínas. Uno de los primeros experimentos significativos en bioinformática fue la aplicación de un programa de búsqueda de similitud de secuencias para la identificación del origen de un gen viral.

Algunas de las aplicaciones más comunes de la Bioinformática son el modelado de procesos biológicos a nivel molecular y la realización inferencias a partir de datos recopilados, de secuencias de ácidos nucleicos de organismos. Una solución bioinformática usualmente involucra los siguientes cuatros pasos: 1) recopilar estadísticas a partir de datos biológicos; 2) construir un modelo computacional del fenómeno biológico; 3) resolver problemas de modelado computacional implementando algoritmos der análisis; y 4) evaluar los algoritmos computacional e interpretar los resultados. (Can, 2014)

Imagen 1 Bioinformática, estructural.

La Bioinformática se limita al análisis de secuencias, análisis de estructuras, análisis de funcionalidad de genes y genomas y de otros productos relacionados, y en ocasiones se considera como Biología molecular computacional. (Xiong, 2006) Imagen 1. Imagen 2.

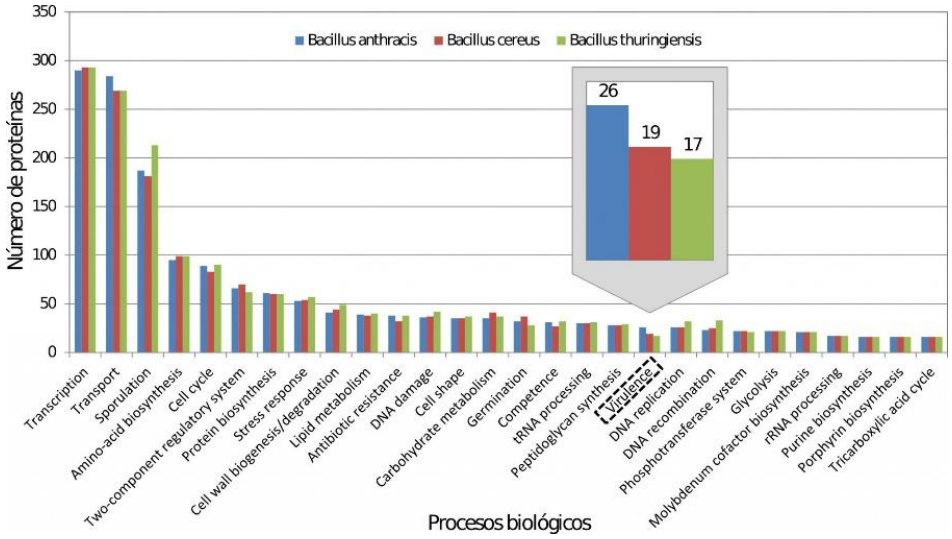


Imagen 2 Bioinformática funcional.

El análisis de secuencias implica inspeccionar de secuencias de ADN, ARN o aminoácidos en busca de pistas sobre su función bilógica, e incluye complicaciones como la identificación de homólogos, alineación de secuencias múltiples, búsqueda de patrones de secuencia y análisis evolutivos. Las estructuras de proteínas son datos tridimensionales y los problemas asociados son la predicción de estructuras (secundaria y terciaria), el análisis de las estructuras de proteínas en busca de pistas sobre la función y la alineación estructural. Los datos de expresión genética generalmente se representan como matrices y el análisis de los datos de microarreglos (gen chips) implica principalmente análisis estadísticos, clasificación y enfoques de agrupamiento. Las redes biológicas, como las redes reguladoras de genes, las vías metabólicas y las redes de interacción proteína-proteína, generalmente se modelan como gráficos y se utilizan enfoques teóricos para resolver problemas asociados, como la construcción y el análisis de redes a gran escala. (Erson-Bensan, 2014)

1.2. Software Bioinformático.

En la actualidad existe una gran variedad de software Bioinformáticos. Entre las herramientas de Software más utilizadas para análisis de genotipos y haplotipos es el Software R, que cuenta con paquetería especializada para este propósito. Por ejemplo: haplo.Stats, que es de uso libre para R, se especializa en análisis estadístico de haplotipos con rasgos y sus covarianzas, cuando la fase de ligamiento es ambiguo. Sus funciones principales son: haplo.em (), haplo.glm (), haplo.score (), y haplo.power (). La función haplo.em permite estimar las frecuencias de haplotipos mediante el algoritmo Expectación-Maximización (EM). La función haplo.glm permite ajustar modelos de regresión logística

empleando las probabilidades de cada par de haplotipos, con incertidumbre como pesos en el modelo. En R hay otros paquetes útiles para análisis de epidemiología genética (genetics, hapassoc y gap) y casi todos emplean una adaptación del programa SNPHAP de Clayton. (DJ., 2016)

Otra herramienta bioinformática es el software Arlequín, cuyo propósito es proporcionar al usuario un conjunto de métodos básicos y pruebas estadísticas, con el fin de extraer información sobre las características genéticas y demográficas de una colección de muestras biológicas de una población.

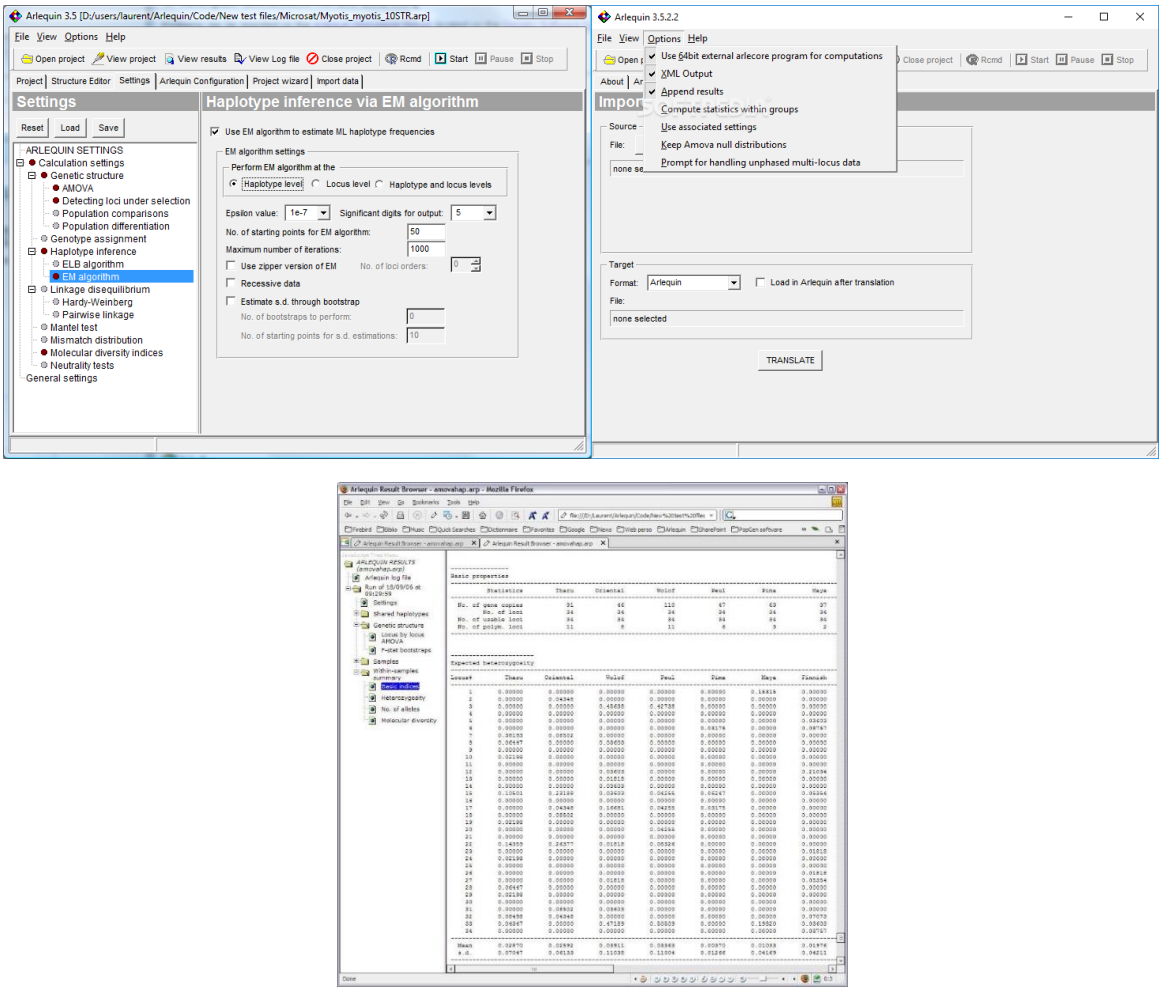


Imagen 3 Software Arlequín, interfaz.

La interfaz gráfica está diseñada para permitir a los usuarios seleccionar rápidamente los diferentes análisis que se desea realizar en sus datos. Pruebas estadísticas son implementadas en Arlequín, para maximizar su potencia. El análisis de Arlequín se divide en dos categorías principales: 1) los métodos intra-poblacionales, y 2) los métodos inter-poblacionales. (Excoffier, 2005)

Una de las herramientas de mayor uso es el software Haploview. Este es un programa de que proporciona el cálculo de las estadísticas de Desequilibrio por ligado (LD), y los patrones de haplotipos de la población, partiendo de genotipos de datos primarios, en una interfaz visualmente atractiva e interactiva.

Imagen 4 Imagen de interfaz gráfica de prueba LD en Haploview.

Haploview está diseñado para simplificar y agilizar el proceso de análisis de haplotipos, proporcionando una interfaz común a varias tareas de análisis. Actualmente Haploview es compatible con las siguientes funcionalidades: cálculo de LD, inferencia de haplotipos, reconstrucción y análisis de bloques de haplotipos, estimación de las frecuencias de

haplotipos de la población, pruebas de asociación de haplotipos con fenotipos, pruebas de permutación para la significancia de la asociación, detección de SNP Taggs, entre otras. Haploview puede importar datos con formato estándar de Mapas de Haplotipos (HapMap) de los proyectos de secuenciación de genomas, así como PLINK, que documenta la relación de parentesco entre los miembros de una población.

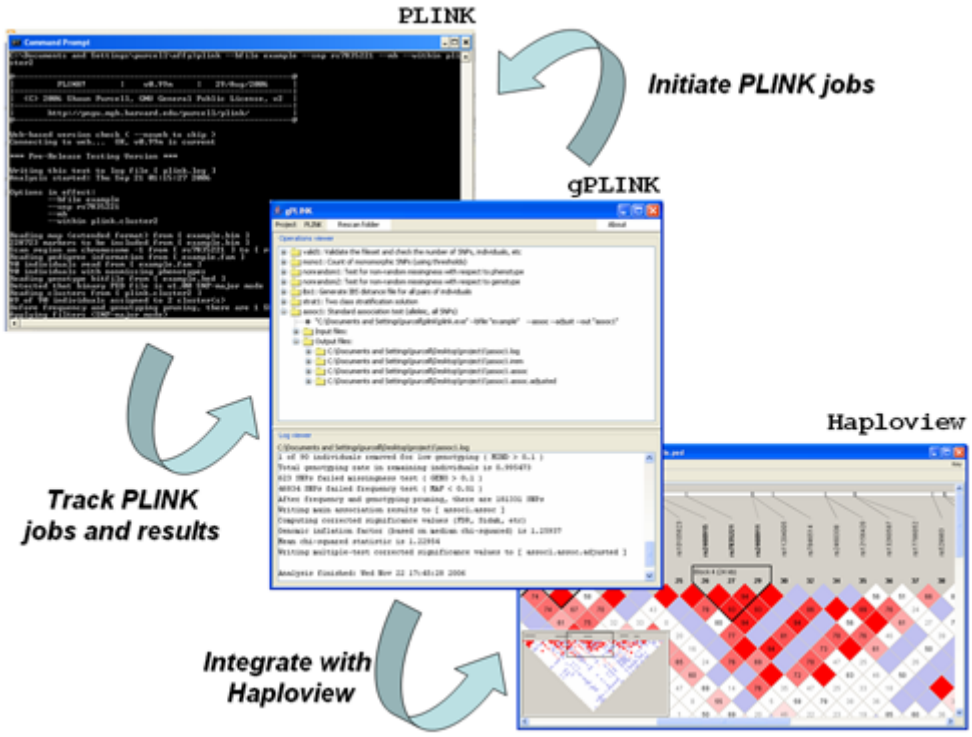


Imagen 5 Trazado PLINK en Haploview.

Haploview es totalmente compatible con los datos del Navegador de Perlegen. Es capaz de analizar miles de SNPs (decenas de miles de personas en el modo de línea de comandos) en miles de individuos.

1.3. Genética y genómica.

1.3.1. Genética.

La Genética es la ciencia que estudia los fenómenos de la herencia y la variación. Estos fenómenos son complejos, y su análisis experimental sólo fue fructífero a partir del momento en que se contó con un marco conceptual adecuado, que fue provisto por el monje austríaco Juan Gregorio Mendel (1822-1884), aunque sus concepciones permanecieron sin uso hasta su redescubrimiento en el año 1900.

La genética ocupa una posición fundamental en el mundo de la biología. Algunos para definirla la definen como “Estudio de la herencia”, pero los fenómenos hereditarios fueron de interés para el humano mucho antes de la biología o genética. La palabra genética proviene de la palabra “gen”, y los genes son el foco del tema, dicho esto la genética es el estudio de los genes. Donde un gen es una molécula de doble hélice llamada ácido desoxirribonucleico, ADN. (Griffiths, 2005). Imagen 6.

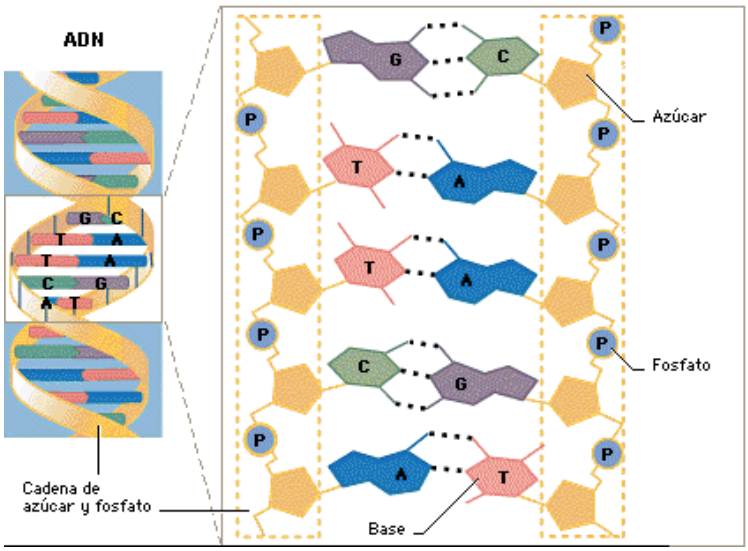


Imagen 6 Acido desoxirribonucleico. Ácido desoxirribonucleico (ADN), material genético de todos los organismos celulares y casi todos los virus. El ADN lleva la información necesaria para dirigir la síntesis de proteínas y la replicación. Se llama síntesis de proteínas a la producción de las proteínas que necesita la célula o el virus para realizar sus actividades y desarrollarse. La replicación es el conjunto de reacciones por medio de las cuales el ADN se copia a sí mismo cada vez que una célula o un virus se reproducen y transmite a la descendencia la información que contiene. En casi todos los organismos celulares el ADN está organizado en forma de cromosomas, situados en el núcleo de la célula.

El trabajo presentado por Mendel en la Sociedad de Ciencias Naturales en Brno (actual República Checa) en 1865 y publicado el año siguiente contiene los postulados teóricos de la Genética, deducidos por Mendel a partir de sus experiencias de hibridación con plantas. Imagen 6.

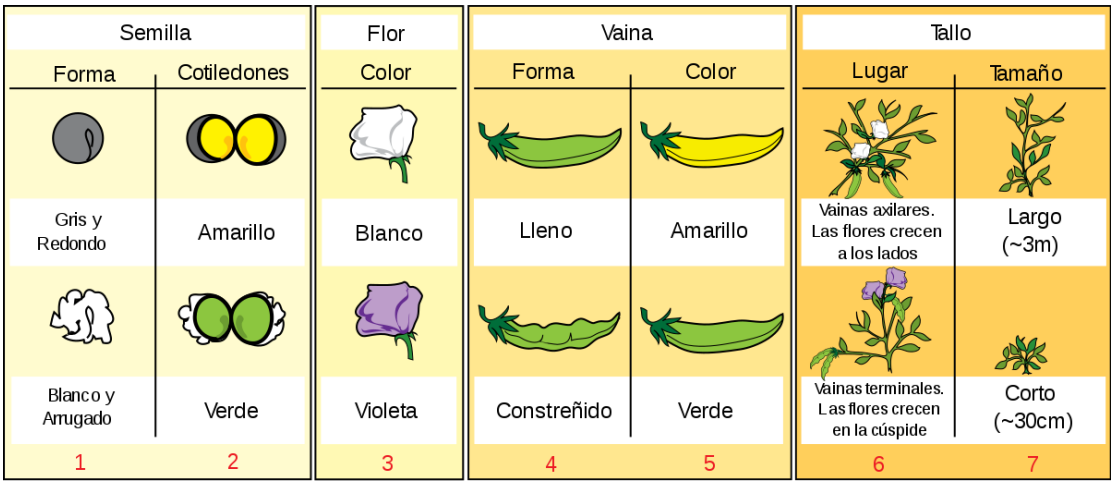


Imagen 7 Hibridación de plantas por Mendel. *a) Elegir una característica de la planta, por ejemplo, color de la semilla del chícharo. b) Se escoge una letra que represente el color de la semilla; en este caso escogeremos la letra A, es sabido que el gen que domina es el amarillo, por lo que quedará representado como AA, lo que significa que tiene dos factores iguales dominantes para esa característica, en tanto el verde - color dominado o rasgo recesivo--, se denotará con la letra v y quedará representado como vv (doble v). Al combinarse estos dos caracteres, las plantas nuevas presentarán un genotipo Av.*

Mendel usó una metodología estadística para establecer reglas cuantitativas para sus resultados de hibridación, efectuados en varios miles de plantas. Para explicar sus resultados, Mendel imaginó "factores" abstractos (décadas después estos factores serían llamados genes) que podían existir en estados alternativos (por ejemplo, un "factor" o gen

para el color verde de la semilla, y un estado alternativo de ese factor, para el color amarillo). Actualmente, los estados alternativos o diferentes de un gen se denominan alelos, término introducido por el genetista W. Johannsen en 1909. En realidad, hoy sabemos que los alelos son todas las variantes que puede presentar un gen por mutación, pero en una simplificación podemos considerar que básicamente hay dos alelos para un factor. Mendel asumió que el origen de las variaciones radicaba en la existencia de "alelos" (normal y mutado, por ejemplo, el color usual y otro color) y que los progenitores contribuían al descendiente con un alelo cada uno. Hoy sabemos que efectivamente nuestros organismos tienen en cada célula sus cromosomas por pares, es decir que nuestros 46 cromosomas son 23 pares, y que por consiguiente tenemos también nuestros genes por pares, condición que se llama diploidía (término introducido recién en 1905 por el citólogo alemán E. Strasburger).

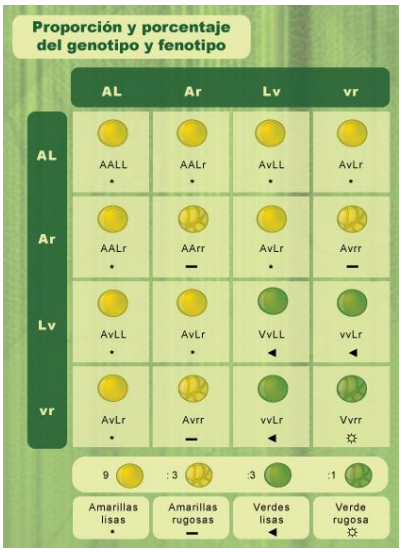


Imagen 8 Estadística Mendel. Para obtener la proporción en la cruza sólo tienes que contar cuantas semillas hay de cada uno de los tipos, por ejemplo, si observas en el cuadro anterior se tienen 9 con semillas amarillas lisas (*), 3 amarillas rugosas (▬), 3 verdes lisas (◀) y 1 verde rugosa (☼),Para el porcentaje, se realiza lo siguiente: 16 posibilidades es el 100%, entonces si tenemos AALL que es 1, podríamos resolverlo con una regla de tres se resuelve multiplicando 1 x 100 = 100 y se divide entre 16, dando como resultado 6.25 %.

Por otra parte, hoy se sabe que efectivamente cada progenitor contribuye con uno solo de cada par de factores o genes, porque las células sexuales (o gametos) sólo poseen un juego de cromosomas en vez de un par de juegos (los gametos humanos tienen 23 cromosomas en vez de los 46 cromosomas de las demás células; son por consiguiente “haploides”, de haplos = mitad, en griego, término también introducido por Strasburger en 1905). Hasta aquí, los postulados puramente hipotéticos de Mendel estaban prediciendo los mecanismos aún no descubiertos de la fertilización y de la meiosis, y, al presumir la existencia de los “factores” (genes) que podían adoptar estados alternativos, predecía los estados de los genes, o normales o mutados. (Desarrollo historico de la genetica humana.)

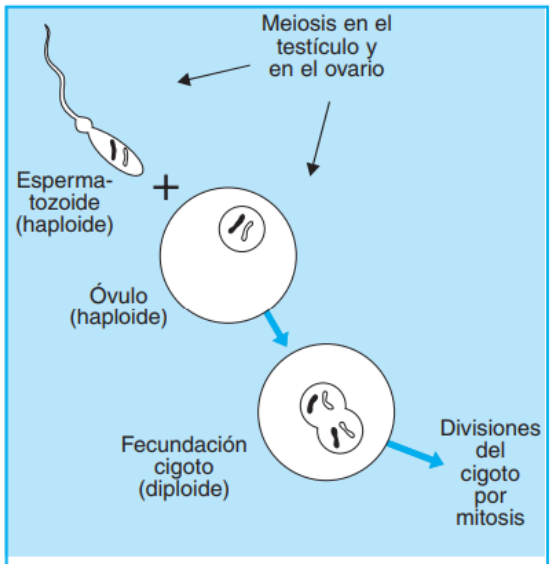


Fig. 1-2. Postulados predichos por Mendel que se comprobaron más tarde por observaciones citológicas. La meiosis (desconocida en tiempo de Mendel) reduce a la mitad el número de cromosomas en los gametos, espermatozoide y óvulo, que son haploides. Al producirse la fecundación, en el cigoto se reconstituye el número de pares de cromosomas, 23 pares en la especie humana (aquí se representan sólo dos pares de cromosomas).

Imagen 9 Postulado de Mendel.

1.3.1.1. Genética de poblaciones.

Es la rama de la biología que proporciona la comprensión más profunda y clara de cómo se produce el cambio evolutivo. La genética de poblaciones es particularmente relevante hoy en día en la búsqueda en expansión para comprender las bases de la variación genética en la susceptibilidad a enfermedades complejas. Muchos de los factores que afectan la frecuencia alélica y las asociaciones entre alelos de genes vinculados se caracterizaron por primera vez en Drosophila y otros organismos modelo, pero los mismos principios se aplican a prácticamente todos los organismos. (Camacho; Camacho., 2002)

Los principios de la genética de poblaciones intentan explicar la diversidad genética en las poblaciones actuales y los cambios en las frecuencias de alelos y genotipos a lo largo del tiempo. Los estudios genéticos poblacionales facilitan la identificación de alelos asociados con el riesgo de enfermedad y proporcionan información sobre el efecto de la intervención médica en la frecuencia poblacional de una enfermedad. Las frecuencias de alelos y genotipos dependen de factores como los patrones de apareamiento, el tamaño y la distribución de la población, la mutación, la migración y la selección. (B.J.B Keats, 2014)

Poco después del redescubrimiento de las leyes de Mendel en 1900, se desarrolló una controversia sobre la relevancia del tipo de variación y transmisión que Mendel caracterizó por la variación suave y continua que los biólogos habían observado y medido en prácticamente todos los organismos. Uno de los argumentos en contra de los genes de Mendel fue que los alelos recesivos pronto se perderían de una población en virtud de su recesión. Godfrey Hardy y Wilhelm Weinberg demostraron de forma independiente la

locura de este argumento, y demostraron que se esperaría que las poblaciones de apareamiento aleatorio retuvieran la variación alélica mediante principios mendelianos simples a menos que alguna otra fuerza actuara sobre la variación.

Imagen 10 Ronald A.Fisher La facilidad matemática que Fisher tenía, le permitió reformular el problema en términos de configurar la muestra en un espacio n-dimensional y mostró que usar la media muestral en lugar de la media poblacional, era equivalente a reducir en uno la dimensionalidad del espacio muestral. De esta manera llegó a un término que después llamó grados de libertad. Esta formulación geométrica del problema lo llevó a derivar la distribución t de Student y, en septiembre de ese año, la envía a W. Gosset, quien ya la había derivado empíricamente, por lo cual se le conoce con su seudónimo.Gosset inmediatamente envió esta demostración a Pearson, sugiriéndole su publicación.

Fue el genetista teórico de la población Ronald Fisher quien desarrolló las matemáticas para mostrar exactamente cuántos genes que actúan juntos podrían producir los grados cuantitativos precisos de semejanza familiar que se observan. Este fue uno de los muchos casos en la historia de la genética de poblaciones. En el que un modelo matemático formal del problema allanó el camino para comprender qué datos empíricos debían recopilarse para probar la nueva conceptualización. Fisher continuó desarrollando, junto con Sewall

Wright y JBS Haldane, gran parte de la teoría para el cambio de frecuencia alélica bajo modelos simples de selección natural. Wright y Fisher desarrollaron la maquinaria teórica necesaria para comprender el complejo proceso de muestreo recurrente que ahora llamamos deriva genética aleatoria. Para 1940, se había desarrollado gran parte de la teoría para la "síntesis moderna" de la evolución darwiniana y la genética mendeliana de transmisión. (Clark, 2001)

El concepto darwiniano de evolución por selección natural es aplicable a entidades que se replican como, por ejemplo, un ser vivo o un algoritmo genético. Los cuatro componentes de este proceso son, Variabilidad, Heredabilidad, Una tasa de alta del crecimiento de la población y Sobrevivencia y reproducción diferencial. El estudio de la evolución ha sido facilitado enormemente por la genética de poblaciones, una disciplina que ha desarrollado herramientas matemáticas que permiten predecir cómo evolucionan los genes en las poblaciones de individuos.

Las poblaciones, como los genes, tienen continuidad de generación en generación y su constitución genética puede cambiar (evolucionar) a través de las generaciones. Aunque la selección natural actúa con mayor intensidad sobre los niveles jerárquicos inferiores (genes, células y, sobre todo, individuos), los cambios evolutivos son más visibles en los niveles superiores (poblaciones, especies y clados). El nivel poblacional (la genética de poblaciones) es, por ahora, el que ha desarrollado las herramientas matemáticas más sofisticadas.

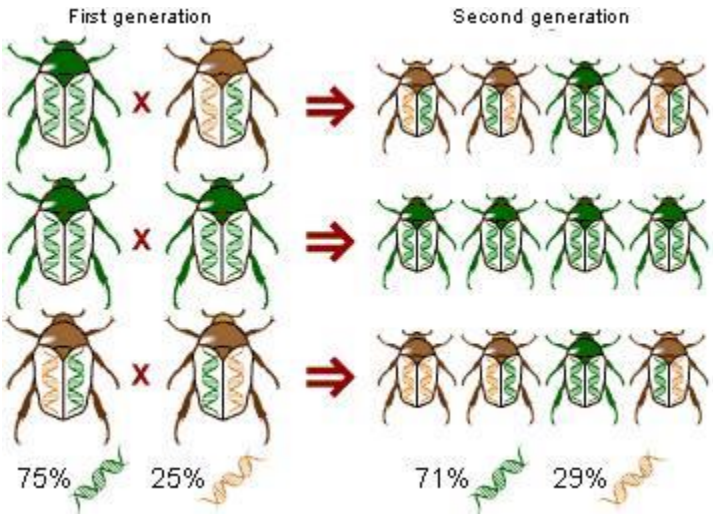


Imagen 11 Genética de poblaciones. Analizando de manera más detallada y precisa en el pool genético o genética de poblaciones se intenta realizar un estudio sobre como afecta el emparejamiento, dentro de lo que se denominarían "genes sociales" sobre la evolución darwininana clásica. En este estudio queda de manifiesto que las reglas de emparejamiento pueden llevar al traste la evolución darwiniana, ya que cuando se ponen las reglas de emparejamiento que van en contra de la mejora de la especie, dichas especies tienden a desaparecer.

Una población es un grupo local de individuos que pertenece a una especie, dentro de la que ocurren apareamientos reales o potencialmente. La población es una entidad genética abierta (que puede intercambiar genes con otras poblaciones de la misma especie), mientras que la especie (y las categorías taxonómicas superiores) es una entidad cerrada (que no puede intercambiar genes con otras entidades). (Camacho., 2002)

El conjunto de informaciones genéticas que llevan todos los miembros de una población, se denomina acervo génico. Para un locus dado, este acervo incluye todos los alelos de dicho gen que están presentes en la población. Puesto que la especie es una entidad genética cerrada, puede decirse que su sino evolutivo está escrito en su acervo, mientras que no ocurre lo mismo con la población.

En genética de poblaciones la atención se centra en la cuantificación de las “frecuencias alélicas y genotípicas” en generaciones sucesivas. Los gametos producidos en una generación dan lugar a los cigotos de la generación siguiente.

Las poblaciones son dinámicas; pueden crecer y expansionarse o disminuir y contraerse mediante cambios en las tasas de nacimiento o mortalidad, o por migración o fusión con otras poblaciones. Esto tiene consecuencias importantes y, con el tiempo, puede dar lugar a cambios en la estructura genética de la población. (Camacho., 2002)

1.3.2. Genómica.

Se define como Genómica a la disciplina que estudia el conjunto del total de los genes, su función y su interacción en el genoma completa de un organismo. Tiene como objetivo catalogar todos los genes que tiene un organismo, estudiar la organización y estructura de cada uno de ellos pero también descubrir la función, los mecanismos implicados en la regulación de la expresión y el modo en que unos genes interaccionan con otros. Desde un punto de vista conceptual, pero también metodológico, los cambios que se están produciendo en esta disciplina son muy importantes. Hasta ahora, con las tecnologías disponibles de Biología Molecular se estudiaban la estructura y función de genes individuales. (Pinto)

La genómica es un vasto campo de la ciencia que involucra a muchas diferentes disciplinas con el enfoque en la plena comprensión cómo funciona el ADN de un organismo (su estructura, función (es) y como son mapeados) y eventualmente pueden ser modificado.

Un conjunto completo de ADN de un organismo, incluyendo todos sus genes, se conoce como un genoma.

La genómica puede confundirse a veces con la genética, pero estos son diferentes. La genética analiza los genes individuales y cómo estos son transmitidos de generación en generación mientras que la genómica es mirando toda la estructura del ADN y el conjunto completo de genes ambos campos están, por supuesto, estrechamente vinculados.

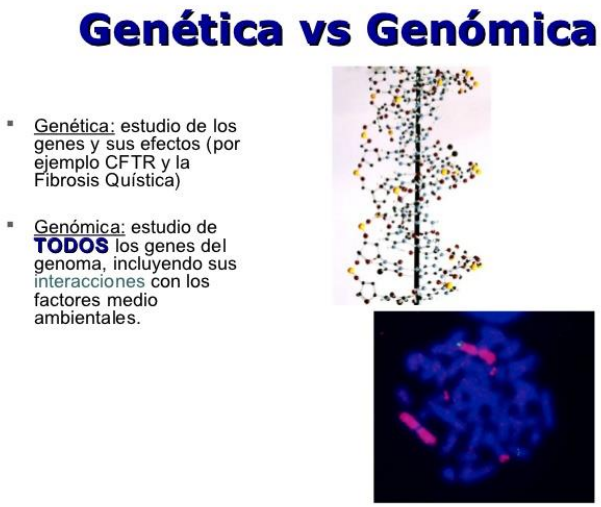


Imagen 12 Genómica Vs genética.

Las aproximaciones de la genómica en cambio permiten el estudio conjunto de los miles de genes, proteínas y metabolitos que constituyen un organismo, así como las complicadas redes de interacciones que entre ellos se establecen en el interior de las células durante su ciclo vital. Como consecuencia de lo anterior, la cantidad de información que se está generando es enorme y por tanto, ha sido necesario el desarrollo en paralelo de potentes herramientas bioinformáticas que permitan almacenar y analizar de forma conveniente los

datos obtenidos. Entre estos estudios de la genómica están la generación de mapas genéticos y físicos de los genomas, generar y ordenas secuencias de genes y genomas, identificar y anotar el conjunto completo de genes que conforman cada genoma, caracterizar la diversidad de secuencias de ADN, proveer los recursos para realizar comparación entre diferentes genomas, establecer bases de datos integrales, basadas en el web, para ofrecer acceso a la información obtenida de todas las investigaciones sobre genomas.

La genómica impacta una parte muy esencial de nuestras vidas y de todos los seres vivos que nos rodean. El genoma humano, es decir, toda su secuencia de ADN, contiene hasta un total de 20,000 a 25,000 genes. (The Genomics Bottleneck, 2015).

Habitualmente la genómica se suele subdividir en dos grandes áreas: La genómica estructural, que se ocupa de la caracterización de la naturaleza física de los genomas, y genómica funcional, cuyo objetivo último es ubicar todos los elementos integrantes de un genoma dentro de una estructura funcional, tanto en el sentido más tradicional de determinar la función de cada una de los elementos componentes de un genoma (las proteínas codificadas, los elementos reguladores, estructurales) como en el sentido más general de determinar la función de que cada uno de estos elementos que desempeña en el funcionamiento global del organismo. (Martinez, 2004)

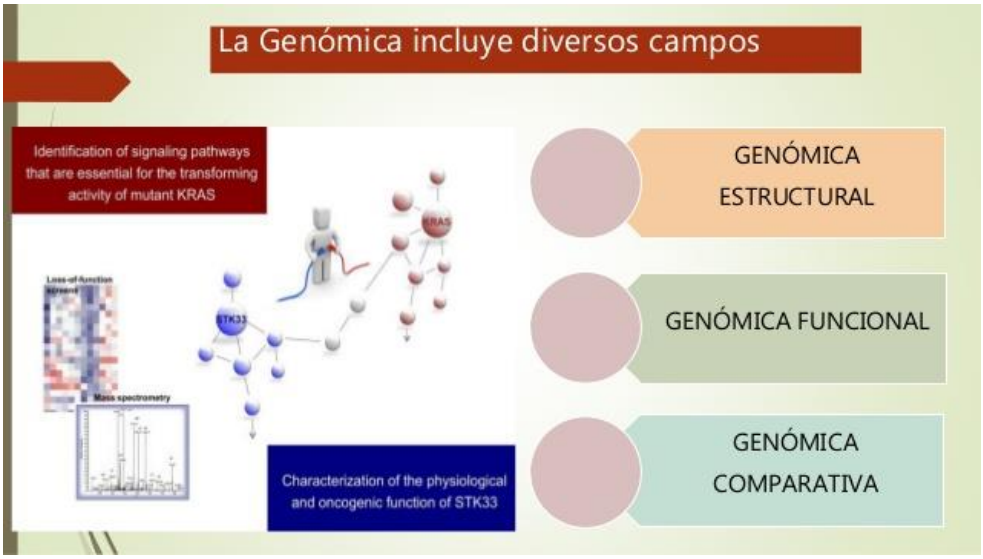


Imagen 13 Campos genómica.

Otras áreas de genómica se encuentran la Genómica Estructural, Genómica Comparativa, Genómica Epigenómica, Farmacogenomica y Metagenómica.

Los distintos métodos de secuenciación son capaces de detectar un máximo de 2Mbp de nucleótidos. El genoma completo de casi todas las especies, y en especial de los mamíferos es demasiado grande (~3X109 Mbp) para ser secuenciado en un conjunto simple de reacciones. Haciendo uso de los secuenciadores actuales, es necesario utilizar estrategias para secuenciar genomas completos. Existen dos estrategias para secuenciar genomas completos, Secuenciación jerárquica y Secuenciación por shotgun. La idea principal de ambas estrategias es fragmentar el genoma completo en millones de piezas, secuenciar las piezas por separado, y posteriormente ensamblar las piezas para formar la secuencia completa del genoma.

1.3.2.1. Genómica de poblaciones.

Es la comparación a gran escala de secuencias de ADN de poblaciones. El campo de la Genómica de Poblaciones inspecciona patrones en el genoma, dentro y entre poblaciones, para hacer inferencias sobre la estructura y evolución del mismo. Un análisis de genómica de poblaciones requiere conjuntos de datos de marcadores o multi-locus de múltiples poblaciones e identifica loci outliers contrastando patrones de divergencia entre regiones genéticas.

Los análisis genómicos de poblaciones requieren conjuntos de datos de múltiples locus de poblaciones e identifican loci no neutrales o atípicos contrastando patrones de divergencia de población entre regiones genéticas. Este enfoque fue propuesto por primera vez por Lewontin & Krakauer (1973), y ahora existen numerosas variaciones sobre este método original (Beaumont 2005, Foll & Gaggiotti 2008, Gompert et al). Quizás el método más comúnmente empleado de estos, particularmente en organismos que no son modelos, es el análisis de valores extremos de FST desarrollado por Beaumont & Nichols (1996).

En el método tradicional de genómica de poblaciones se comprende el muestreo de gran cantidad de individuos, genotipado de los individuos muestreados, para muchos locus independientes, identificación estadística de valores extremos (otliers), estimar los parámetros estadísticos o demográficos en una población grande, removiendo los valores extremos, o alternativamente, estudiando los locus de valor extremo para inferir mecanismos potenciales de selección, que rigen la estructura genética de la población.

Los avances recientes en métodos computacionales y biología molecular (incluida la secuenciación de próxima generación) permiten que se investiguen los patrones de divergencia genómica a escalas previamente inalcanzables.

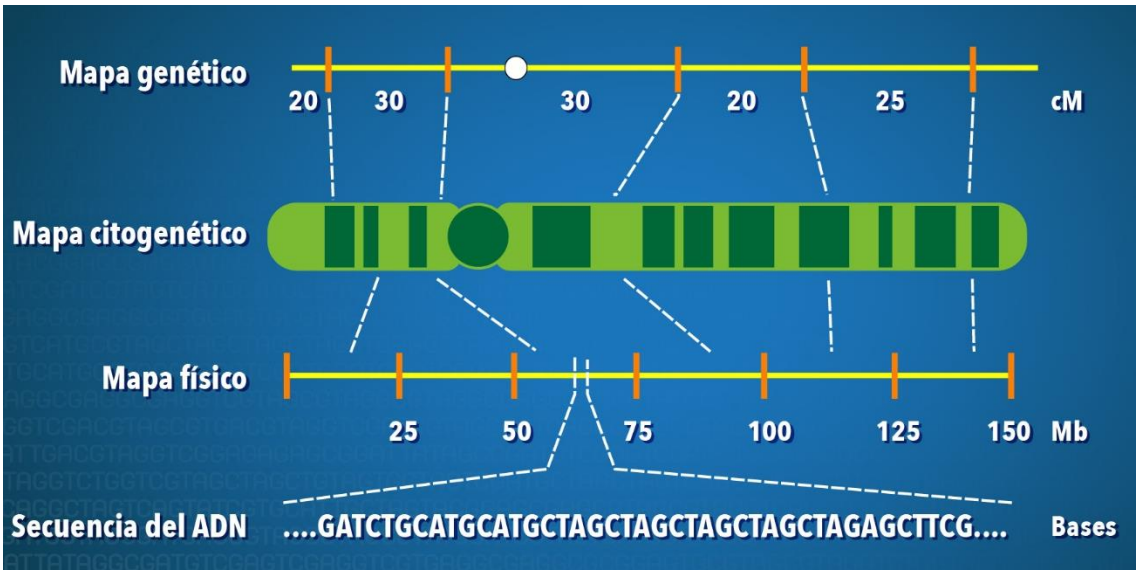


Imagen 14 Mapeo genético.

1.4. Marcadores genéticos.

Se es posible denominar a la expresión fenotípica observada como marcador génico, es decir, aquel que establece una relación inequívoca entre ella y la presencia del gen que la controla (Hattemer et al. 1993). El análisis genético de la variación fenotípica es, por lo tanto, la única comprobación aceptable para poder referirse a genes o genotipos y locus o loci génicos. Resulta obvio, entonces, que en estudios de genética poblacional y evaluaciones de genética cuantitativa sea siempre recomendable, aunque no imprescindible, trabajar con marcadores génicos.

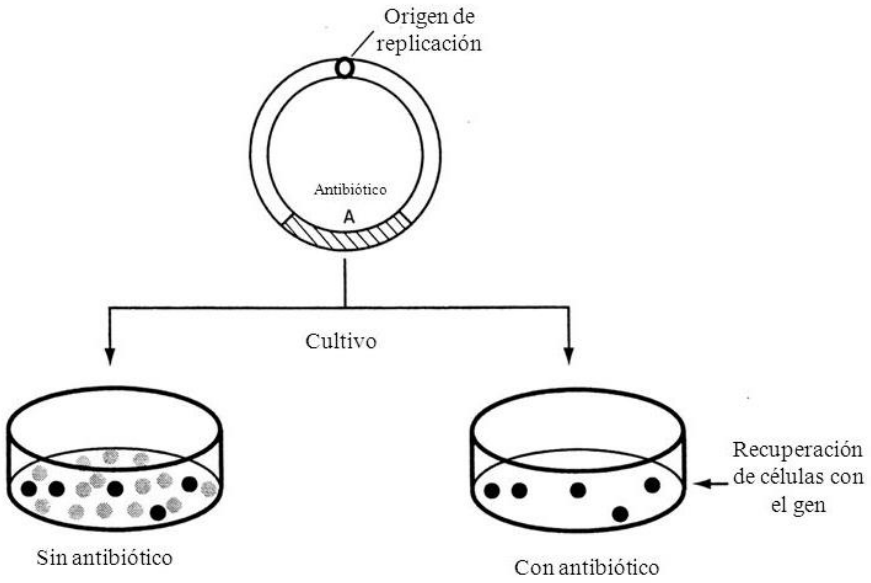


Imagen 15 Marcador genético.

Los primeros marcadores génicos fueron morfológicos y corresponden a los determinados por Mendel en sus históricos y notables estudios con material segregante de Pisum sp. y Phaseolus sp., a partir de los cuales formuló las leyes que aún hoy rigen en la genética (Mendel 1866). (Gallo, 2006)

Por lo que a un marcador genético se le conoce como un segmento de ADN con una ubicación física conocida en un cromosoma. Los segmentos de ADN que se encuentran cerca en un cromosoma tienden a heredarse juntos. Los marcadores genéticos se utilizan para rastrear la herencia de un gen cercano que aún no ha sido identificado, pero cuya localización aproximada es conocida. El marcador genético en sí puede ser parte de un gen o puede no tener ninguna función conocida.

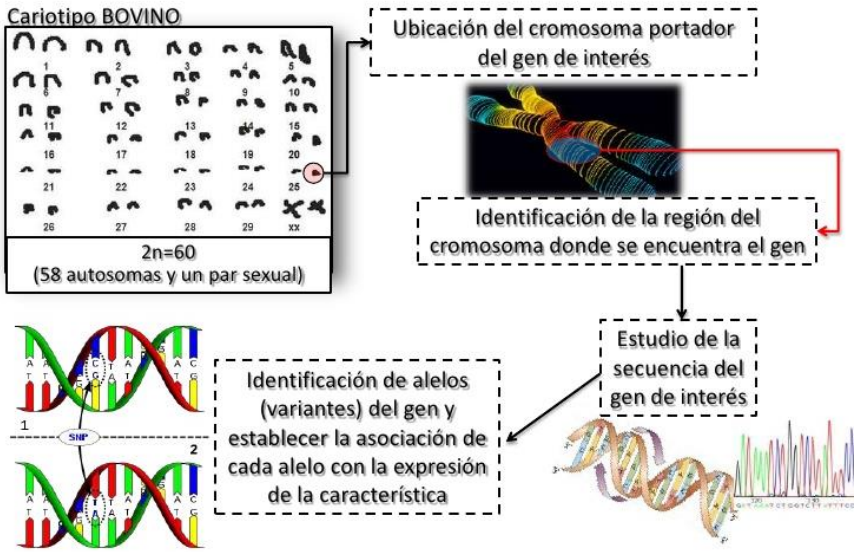


Imagen 16 Identificación de marcador genetico.

Un marcador genético, puede identificarse como cualquier alteración en una secuencia de ácidos nucleicos u otro rasgo genético que pueda detectarse fácilmente y usarse para identificar individuos, poblaciones o especies, o para identificar genes involucrados en enfermedades hereditarias. Marcadores genéticos consisten principalmente de polimorfismos, que son variaciones genéticas discontinuas que dividen a los individuos de una población en formas distintas (p. ej., tipo de sangre, color de cabello, color de ojos). Los marcadores genéticos desempeñan un papel clave en el mapeo genético, específicamente en la identificación de las posiciones de los diferentes alelos que se encuentran cerca uno del otro en el mismo cromosoma y tienden a heredarse juntos. Dichos grupos de enlace pueden usarse para identificar genes desconocidos que influyen en el riesgo de enfermedad. Los avances tecnológicos, especialmente en la secuenciación de ADN, han aumentado considerablemente el catálogo de sitios en el estudio del genoma humano.

Los múltiples tipos de polimorfismos sirven como marcadores genéticos, incluidos los polimorfismos de nucleótido simple (SNP), los polimorfismos de longitud de secuencia simple (SSLP) y los polimorfismos de longitud de fragmentos de restricción (RFLP). Los SSLP incluyen secuencias de repetición, variaciones conocidas como minisatélites (número variable de repeticiones en tándem, o VNTR) y microsatélites (repeticiones en tándem simples, STR). Las inserciones son otro ejemplo de un marcador genético.

Los marcadores genéticos ahora se usan ampliamente en muchos campos de la biología de la población y pueden analizarse utilizando varios enfoques. Entre ellos, los métodos multivariados, como el análisis de componentes principales (PCA), están obligados a jugar un papel importante porque pueden resumir la variabilidad genética sin hacer suposiciones sólidas sobre un modelo de evolución: no se basan en el equilibrio de Hardy-Weinberg, ni suponen la ausencia de desequilibrio de ligamiento. Recientemente, los métodos multivariables han demostrado ser útiles para evaluar la estructura genética de consenso entre un conjunto de marcadores genéticos, así como para investigar el patrón espacial de la variabilidad genética. Sin embargo, los métodos multivariables actualmente disponibles en el software de genética de poblaciones son muy restringidos, a pesar del número bastante grande de estos programas. (Jombart, 2008).

Los datos de marcadores genéticos básicos son genotipos obtenidos para un conjunto de marcadores, cada alelo está codificado por una cadena de caracteres (Warnes, 2003). Para usar métodos estadísticos, dicha información no se puede usar directamente, y necesita ser recodificada numéricamente en una matriz de frecuencias alélicas.

1.4.1. Marcadores SNP

Los marcadores SNP o Polimorfismo de Nucleótido Simple ("SNP Single Nucleotide Polimorphism" por sus siglas en Ingles) afectan solo a uno de los componentes básicos del cromosoma: adenina (A), guanina (G), timina (T) o citosina (C) en un segmento de ADN. Por ejemplo, en una ubicación genómica con la secuencia ACCTGA en la mayoría de los individuos, algunas personas pueden contener ACGTGA en su lugar. La tercera posición en este ejemplo se consideraría un SNP, ya que existe la posibilidad de que ocurra un alelo C o un alelo G en la posición variable. Debido a que cada individuo hereda una copia del ADN de cada padre, cada persona tiene dos copias complementarias del ADN. Como resultado, en el ejemplo anterior, tres genotipos son posibles: CC homocigotos (dos copias del alelo C en la posición variable), CT heterocigotos (un alelo C y uno T) y TT homocigotos (alelos T dos). Los tres grupos de genotipos pueden usarse como categorías de "exposición" para evaluar asociaciones con un resultado de interés en un entorno de epidemiología genética. Si se identifica una asociación de este tipo, se puede investigar la región genómica marcada para identificar la secuencia de ADN particular en esa región que tiene un efecto biológico directo en el resultado de interés. (Britanica, 2018). Pueden actuar como marcadores biológicos, ayudando a los científicos a localizar genes asociados con enfermedades. Cuando ocurren SNPs dentro de un gen o en una región reguladora cerca de un gen, pueden jugar un papel más directo en la enfermedad al afectar la función del gen.

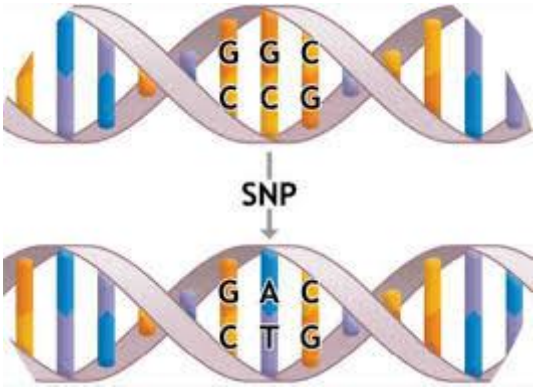


Imagen 17 Marcador SNP.

La mayoría de los SNP no tienen ningún efecto sobre la salud o el desarrollo. Algunas de estas diferencias genéticas, sin embargo, han demostrado ser muy importantes en el estudio de la salud humana. Los investigadores han encontrado SNP que pueden ayudar a predecir la respuesta de un individuo a ciertos medicamentos, la susceptibilidad a factores ambientales como las toxinas y el riesgo de desarrollar enfermedades particulares. Los SNP también se pueden usar para rastrear la herencia de genes de enfermedades dentro de las familias. Los estudios futuros trabajarán para identificar los SNP asociados con enfermedades complejas como las enfermedades cardíacas, la diabetes y el cáncer. (nlm, 2019)

1.5. Parámetros genéticos basados en SNPs.

1.5.1. Frecuencias alélicas

Del examen fenotípico de los individuos de una población se deduce la existencia de variaciones fenotípicas, muchas de las cuales provienen de la diversidad genética

subyacente. Cuando existe correspondencia biunívoca e inequívoca entre fenotipo y genotipo, podemos calcular las "frecuencias alélicas". Por ello, primero es necesario conocer la naturaleza genética de los caracteres variables, mediante el análisis de los resultados de determinados apareamientos, y entonces podremos definir la población por el número relativo de genotipos de cada clase en los adultos (frecuencias genotípicas) y el número relativo de alelos de cada tipo en los gametos (frecuencias alélicas).

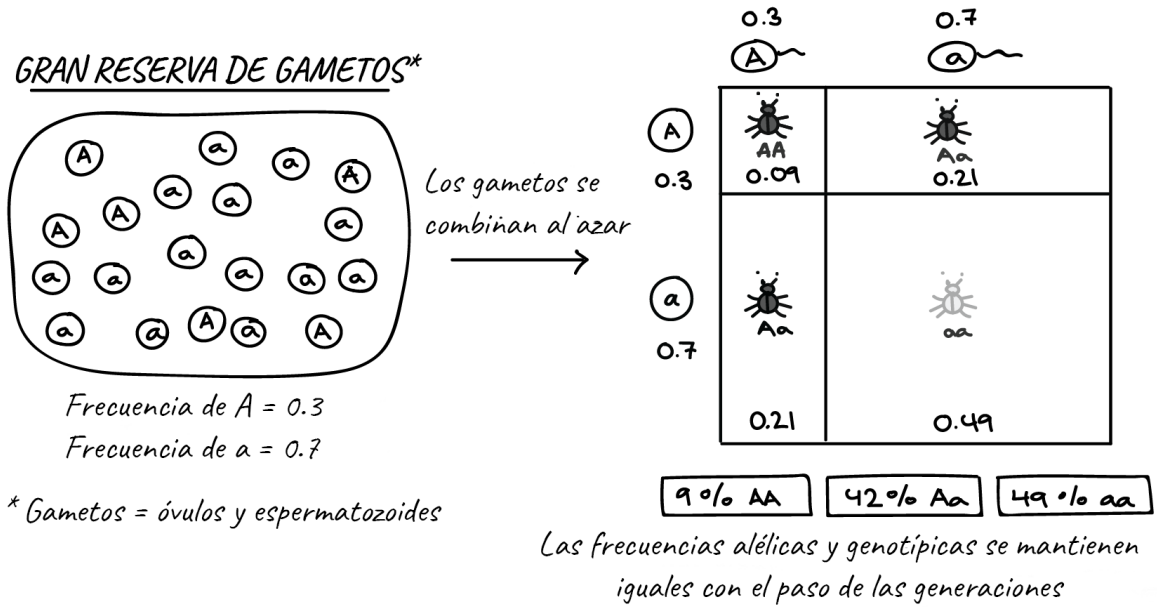


Imagen 18 Ejemplo de cálculo de frecuencias alélicas.

Supongamos un gen con los alelos A y a, y sean nAA, nAa y naa los números de individuos con los genotipos AA, Aa y aa, respectivamente, de modo que nAA+ nAa+naa=N, siendo N el nº total de individuos de la población. Si representamos por X, Y y Z las proporciones de los genotipos AA, Aa y aa en la población, las frecuencias genotípicas serán:

$$X_{AA}=\frac{n_{AA}}{N}, \quad Y_{Aa}=\frac{n_{Aa}}{N} \text{ y } Z_{aa}=\frac{n_{aa}}{N},$$

De forma que $X_{AA}+Y_{Aa}+Z_{aa}=1.$

Las frecuencias alélicas (denominemos p a la frecuencia de A y q a la de a) se calculan a partir del número de individuos de cada genotipo:

$$p=\frac{2n_{AA}+n_{Aa}}{2N}=\frac{n_{AA}+\frac{1}{2}n_{Aa}}{N}, \text{ y } q=\frac{2n_{aa}+n_{Aa}}{2N}=\frac{n_{aa}+\frac{1}{2}n_{Aa}}{N}$$

O bien a partir de las frecuencias genotípicas:

$$p=\frac{n_{AA}+\frac{1}{2}n_{Aa}}{N}=\frac{n_{AA}}{N}+\frac{1}{2}\frac{n_{Aa}}{N}=X_{AA}+\frac{1}{2}Y_{Aa}$$
$$q=\frac{n_{aa}+\frac{1}{2}n_{Aa}}{N}=\frac{n_{aa}}{N}+\frac{1}{2}\frac{n_{Aa}}{N}=Z_{aa}+\frac{1}{2}Y_{Aa}$$

Como p+q=1, basta calcular p para conocer q: q=1-p. Es decir, “la frecuencia del alelo A se determina sumando la mitad de la frecuencia de heterocigotos a la frecuencia de homocigotos AA”, y “la frecuencia del alelo a es igual a la suma de la mitad de la frecuencia de heterocigotos y la proporción de homocigotos aa”. A nivel genético, el cambio evolutivo en una población puede describirse como cambio en las frecuencias de los diferentes alelos y en las frecuencias genotípicas. Si todos los individuos de una población son genéticamente idénticos para un cierto locus, es decir, son homocigotos para el mismo alelo (p=1), no puede haber evolución en ese locus, ya que las frecuencias alélicas no pueden cambiar de una generación a otra. Por el contrario, si en otra población hay dos alelos en ese locus, sí puede haber cambio evolutivo en ese locus porque la frecuencia de un alelo puede

incrementar a expensas de la del otro. Esto supone que cuanto mayor sea el número de loci variables y más alelos haya en cada locus variable, mayor será la probabilidad de que las frecuencias alélicas cambien en algún locus.

1.5.2. Proporciones polimórficas

Los polimorfismos genéticos son variantes del genoma que aparecen por mutaciones en algunos individuos, se transmiten a la descendencia y adquieren cierta frecuencia en la población tras múltiples generaciones. Los polimorfismos son la base de la evolución y los que se consolidan, bien pueden ser silentes o proporcionar ventajas a los individuos, aunque también puede contribuir a causar enfermedades. Se conocen muchas enfermedades determinadas genéticamente por mutaciones ya que los portadores de la variante suelen manifestar la enfermedad con una alta probabilidad. Estas variantes suelen ser de baja frecuencia en la población general.

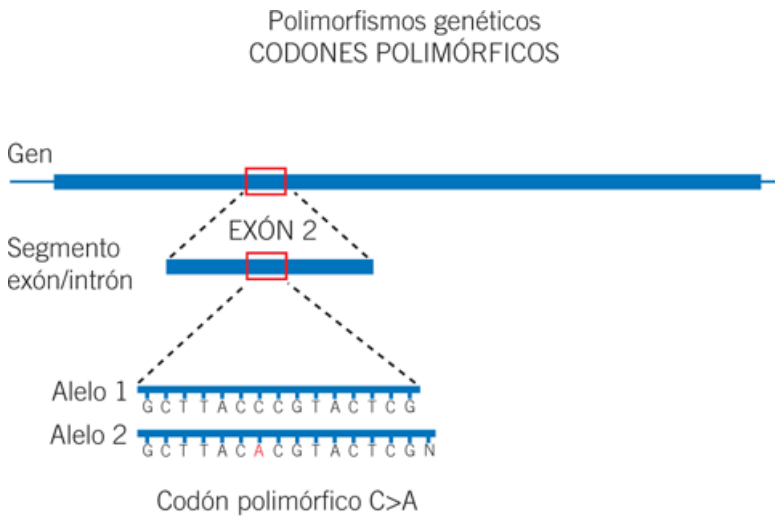


Imagen 19 Polimorfismos genéticos. Un gen está representado como una sección de cromosoma, que equivale a un segmento de ADN con una secuencia determinada, con la información necesaria para crear una proteína específica. Los genes de un organismo contienen toda la información de manera precisa acerca de cada aspecto y proceso de la

formación y desarrollo de un individuo; si bien, los genes no pueden dictar su comportamiento, por lo que hay que considerar que el ambiente desempeña un papel muy importante en cómo estos genes se expresan a lo largo de la vida del organismo.

Los polimorfismos más frecuentes son cambios de una única base. A éstos se les llama polimorfismos de un único nucleótido (single nucleotide polymorphism [SNP). Por ejemplo, en el gen de la apolipoproteína E (ApoE) se han descrito varios polimorfismos frecuentes que consisten en cambios de una única base. Uno de ellos, denominado ApoE * -4, resulta en un cambio en el aminoácido cisteína de la posición 112 por una arginina. Esta variante se asocia con la enfermedad de Alzheimer7. Otros polimorfismos son repeticiones, en un número variable de veces, de una secuencia corta (variable number tandem repeat [VNTR]). Por ejemplo, los individuos afectados por ataxia de Fredreich, una enfermedad autosómica recesiva, son portadores de variantes en el gen frataxin con un número elevado repeticiones del triplete GAA en el primer intrón. Los individuos normales suelen tener menos de 40 repeticiones, mientras que los afectados tienen entre 100 y 1.700 repeticiones.

Los polimorfismos se deben a lesiones o inserciones de secuencias cortas de nucleótidos. El cambio de un único nucleótido, si ocurre en una zona codificante, puede provocar un cambio de aminoácido en la proteína resultante, y ello puede resultar en una modificación de su actividad o función. Los cambios también pueden ocurrir en zonas del promotor de un gen y modificar su expresión. Estas zonas promotoras modulan el proceso de transcripción del ADN en ARN (la transcripción es el primer paso de la decodificación de un gen a una proteína). Lo mismo puede ocurrir si el cambio se produce en un intrón. Aunque los intrones no se traducen a proteína, cambios en su estructura pueden modular la

expresión del gen. Otras veces, probablemente la mayoría, los cambios son silentes y no tienen repercusiones funcionales. Mientras que sólo estudios moleculares específicos pueden poner de manifiesto si los polimorfismos son funcionales, los estudios epidemiológicos son fundamentales para valorar si hay efectos en la salud de la población.

Un polimorfismo se caracteriza porque diferentes individuos presentan distintos nucleótidos o variantes en una posición concreta del genoma, que se denomina locus. A cada posible variante se le denomina alelo. Si se trata de un SNP, normalmente serán 2 los posibles alelos en un locus: por ejemplo, el cambio de T por C (T > C). Si el locus corresponde a un cromosoma autosómico (del 1 al 22), cada individuo es portador de 2 alelos, uno en cada copia del cromosoma, que se heredan del padre y madre de manera independiente. La pareja de alelos observada en un individuo se denomina genotipo y, para el locus T > C del ejemplo, las 3 posibilidades de parejas de alelos son: TT, TC y CC. Los individuos con los 2 alelos idénticos, sean TT o CC, se denominan homocigotos y los que tienen diferentes alelos (TC), heterocigotos. En general se considera variante al alelo menos frecuente, pero esto puede diferir de una población a otra.

La descripción estadística de un polimorfismo consiste, en primer lugar, en estimar la prevalencia en la población de cada alelo y de cada genotipo posible, lo que en nomenclatura genética se denomina estimar las frecuencias alélicas y genotípicas, respectivamente. En general, las técnicas de laboratorio permiten determinar el genotipo de cada individuo. Las frecuencias genotípicas, por tanto, se estiman directamente calculando la proporción de individuos con cada genotipo. Para estimar las frecuencias alélicas simplemente se duplica la muestra tomando como unidad de observación el

cromosoma (cada individuo contribuye con 2 cromosomas) y se calcula la proporción de cada alelo. (Raquel & GUINO, 2005)

Modelo[a]	Genotipo[b]	Controles		Casos		OR	IC del 95%	p[c]	ΔCo[d]
		N	%	N	%				
Co	TT	210	65,0	225	62,2	1		0,07	
	TC	104	32,2	114	31,5	1,02	0,74-1,42		
	CC	9	2,8	23	6,4	2,38	1,09-5,22		
Do	TT	210	65,0	225	62,1	1		0,43	0,034
	TC-CC	113	34,9	137	37,8	1,13	0,83-1,55		
Re	TT-TC	314	97,2	339	93,6	1		0,024	0,89
	CC	9	2,79	23	6,3	2,37	1,09-5,14		
Ad						1,21	0,93-1,57	0,14	0,08

Imagen 20 Análisis del riesgo de un polimorfismo en función del modelo de herencia. Se presentan los 4 modelos principales de riesgo posibles. El modelo codominante muestra que sólo los individuos CC tienen un riesgo aumentado. Los heterocigotos TC tienen un riesgo similar a los homocigotos TT, lo que sugiere que el modelo recesivo es el adecuado.

1.5.3. Diversidad de nucleótidos

Es el cambio aleatorio que sufren las frecuencias de alelos, de una generación a otra, debido al apareamiento aleatorio entre individuos de la población.

Normalmente se da una pérdida de los alelos menos frecuentes y una fijación (frecuencia próxima al 100%) de los más frecuentes, resultando una reducción en la diversidad genética de la población.

Los efectos de deriva genética se acentúan en poblaciones de tamaño pequeño provocando cambios que no son necesariamente adaptativos (puede ocurrir en efectos de cuello de botella y efectos fundadores).

La deriva genética tiende a formar una población homocigótica, es decir, tiende a eliminar los genotipos heterocigotos.

Debido a que en cada población pueden ser distintos los alelos que se pierden y se fijan, la deriva genética hace que dos poblaciones de la misma especie tiendan a diferenciarse genéticamente. La deriva genética es estimada a partir de la varianza entre las frecuencias alélicas.

Supongamos que observamos un número grande de poblaciones por separado, cada una de tamaño Ny frecuencias alélicas p y q. Después de una generación de apareamiento al azar, la deriva genética será:

$$s_{\sigma^2} = \frac{pq}{2N}$$

1.5.4. Heterosigocidad

La heterocigosidad media es una medida de la variación de los genotipos. Se estima calculando la frecuencia de los heterocigotos para cada locus y dividido por el total de loci.

Heterocigosidad se define como la medida de la variación genética de una población respecto a un locus particular. Se define como la frecuencia de heterocigotos para ese locus.

Las células diploides, por ejemplo las células somáticas humanas, contienen 2 copias del genoma, cada una de ellas procede de un progenitor. La posesión de 2 alelos idénticos en una posición determinada se denomina homocigosidad, si los alelos son diferentes se nombra como heterocigosidad. La pérdida de heterocigosidad puede definirse como el

fenómeno por el cual un locus (posición precisa de un gen en un cromosoma) pierde una de las copias de un gen, por deleción u otro mecanismo. Este proceso se llama también LOH por sus siglas en inglés (Loss Of Heterozygosity).

Definición de las mediciones jerárquicas de heterosigosidad:

H_I = Heterosigosidad promedio observada por individuo dentro de las subpoblaciones.

H_S = Heterosigosidad esperada entre subpoblaciones con apareamiento aleatorio = 2piqi.

H_T = Heterosigodsidad esperada en la población total con apareamiento aleatorio = 2pq.

1.5.5. Coeficiente de endogamia

Las tres estadísticas-F jerárquicas son definidas como sigue:

F_{IS} = Coeficiente de endogamia (Inbreeding coefficient): es la reducción promedio de H0 en un individuo debido al apareamiento no aleatorio dentro de una subpoblación.

$$F_{IS} = \frac{H_S - H_I}{H_S}$$

• F_{IS} es una medida del grado de endogamia (consanguinidad) dentro de una sub población.

• Varia desde -1 (todos los individuos son heterocigotos) a +1 (no heterocigotos observados).

1.5.6. Índice de fijación

El índice de fijación, representado como F_{st}, es un valor estadístico empleado para evaluar el nivel de diferenciación genética entre poblaciones.

Se trata de un término acuñado por el genetista estadounidense Sewall Wright y que forma parte de lo que se denominan estadísticos F de Wright. El índice de fijación es un componente generado por el balance que se establece entre el flujo genético (por ejemplo debido a migraciones) y la deriva genética, que en poblaciones aisladas genéticamente, (ya sea por creencias religiosas, políticas o aislamiento geográfico) pueden tener un efecto importante en la estructura genética de los individuos que las componen. Teniendo esto último en cuenta, también se utiliza como definición del índice de fijación lo siguiente: "probabilidad de poseer dos alelos idénticos por descendencia (es decir, heredados o procedentes de un alelo común ancestral) en un determinado locus."

Como ya se ha comentado anteriormente, el valor de F_{st} se utiliza habitualmente en estudios de genética de poblaciones.

Un ejemplo reciente, y que representa muy bien lo que más arriba hemos denominado aislamiento genético, lo encontramos en un trabajo publicado en la revista Nature Genetics. En este trabajo, realizado en la población de la isla de Cerdeña, se analizó el valor del índice de fijación con el objetivo de medir la diferenciación de esta población con otras procedentes de Europa, a partir de las cuales se pobló la isla. Los resultados mostraron como existe una clara diferenciación genética entre la población sarda y las distintas poblaciones europeas de procedencia, que puede explicarse por el efecto de la deriva

genética y selección natural que sufrió la población de la isla tras el asentamiento inicial. (Índice de fijación - Wikipedia, la enciclopedia libre)

F_{ST} = Índice de fijación (fixation index): es la reducción promedio de

H_0 en una subpoblación (en relación a la población total) debido a deriva genética entre subpoblaciones

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

• F_{ST} es una medida del grado de diferenciación genética entre subpoblaciones .

• Varia desde 0 (no existe diferenciación) a 1 (existe total diferenciación, lo que significa. que las subpoblaciones se han fijado para diferente alelos.

1.5.7. Índice global de fijación

F_{ST} para varios organismos, F_{IT} = Índice global de fijación (overall fixation index): es la reducción promedio de H_0 en un individuo en relación población total

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

• F_{IT} combina las contribuciones entre apareamiento no aleatorio dentro “demes” (F_{IS}) y efectos de deriva aleatoria en todos los “demes” (F_{ST}).

1.5.8. Desequilibrio por Ligamento

La expresión “desequilibrio de ligamiento”, abreviada como LD, es una medida de la correlación de los alelos segregados en un locus, comparado con otro locus. El LD es una de las más desafortunadas expresiones usadas en genética cuantitativa, puesto que puede no haber ligamiento entre genes y existir desequilibrio de ligamiento, o por el contrario genes ligados pueden no presentar este desequilibrio de ligamiento. La forma correcta de expresar la relación entre genes o alelos es decir si hay o no independencia estadística entre ellos. Supongamos que un carácter está regulado por dos loci con dos alelos cada uno, (A,a) y (B,b) en cromosomas diferentes. Supongamos que podemos identificar a los nueve tipos de individuos que existirían en nuestra población: AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb, aabb. Si los alelos fueran independientes estadísticamente, se podría calcular las frecuencias de los genotipos conociendo las frecuencias génicas.

Definición: Se llama desequilibrio de ligamiento a la diferencia entre la frecuencia observada de un gameto y la que debería tener si estuviera en equilibrio (esto es, en condiciones de independencia estadística).

Hay varias consecuencias para la mejora genética:

- Las asociaciones entre marcadores moleculares (SNPs, microsatélites, etc.) y genes (tanto QTLs como genes menores) se irán deshaciendo de generación en generación. Esto implica que no se pueden aplicar las asociaciones encontradas en una generación ni a datos antiguos ni a datos nuevos después de unas generaciones.

- Puede haber marcadores no ligados a genes cuantitativos que sin embargo estén en desequilibrio con ellos, y seleccionando el marcador se seleccione también el gen de interés, aunque estas asociaciones desaparezcan rápidamente en las siguientes generaciones. Se trata simplemente de que no estén en estado de independencia estadística.

- El desequilibrio de ligamiento; esto es, la falta de dependencia estadística, se genera de varias formas, una de ellas es la selección, puesto que al seleccionar estamos alterando las frecuencias de los gametos y favoreciendo a unas frente a otras, con lo que la independencia estadística se pierde.

- Cuando hay desequilibrio de ligamiento, por ejemplo, cuando se cruzan dos razas, hay caracteres que aparecen como más relacionados de lo que están, y su asociación se va perdiendo con el tiempo. Hay que tener cuidado al estimar correlaciones genéticas en poblaciones en las que hay desequilibrio de ligamiento, por ejemplo en poblaciones seleccionadas.

- Asumimos que en el modelo infinitesimal los caracteres se distribuyen de forma Normal porque están controlados por muchos genes independientes de pequeño efecto cada uno. La selección produce desequilibrio de ligamiento (dependencia estadística) como acabamos de decir, por lo que no se cumple una de las condiciones del teorema central del límite y no se puede invocar a la normalidad en este caso. Aunque existe la sensación de que este efecto no es importante, depende de las circunstancias (tamaño efectivo de la población, intensidad de la selección, etc.).

- Otra consecuencia es que la varianza aditiva total ya no es la suma de las de cada gen, sino que hay que contar las covarianzas entre genes; esto es particularmente importante en el caso de que la población esté seleccionada, porque se ha producido desequilibrio de ligamiento (dependencia estadística) por la selección y estas covarianzas ya no son nulas. De hecho son negativas, aunque parezca contraintuitivo (da la impresión de que al seleccionar, los genes que van en la misma dirección deberían aumentar todos su frecuencia y las covarianzas ser positivas), y la varianza aditiva disminuye; es lo que se conoce como "Efecto Bulmer". (ACTEON)

1.5.9. Frecuencias de Hardy-Weinberg

En genética de poblaciones, el principio de Hardy-Weinberg (PHW) (también equilibrio de Hardy-Weinberg, ley de Hardy-Weinberg o caso de Hardy-Weinberg) establece que la composición genética de una población permanece en equilibrio mientras no actúe la selecció natural ni ningún otro factor y no se produzca ninguna mutación. Es decir, la herencia mendeliana, por sí misma, no engendra cambio evolutivo. Recibe su nombre del matemático inglés G. H. Hardy y del médico alemán Wilhelm Weinberg, que establecieron el teorema independientemente en 1908.

En el lenguaje de la genética de poblaciones, la ley de Hardy-Weinberg afirma que, bajo ciertas condiciones, tras una generación de apareamiento al azar, las frecuencias de los genotipos de un locus individual se fijarán en un valor de equilibrio particular. También especifica que esas frecuencias de equilibrio se pueden representar como una función sencilla de las frecuencias alélicas en ese locus. En el caso más sencillo, con un locus con dos alelos A y a, con frecuencias alélicas de p y q respectivamente, el PHW predice que la frecuencia genotípica para el homocigoto dominante AA es p2, la del heterocigoto Aa es 2pq y la del homocigoto recesivo aa,

es q2. El principio de Hardy-Weinberg es una expresión de la noción de una población que está en "equilibrio genético", y es un principio básico de la genética de poblaciones. (Bodmer, 1981)

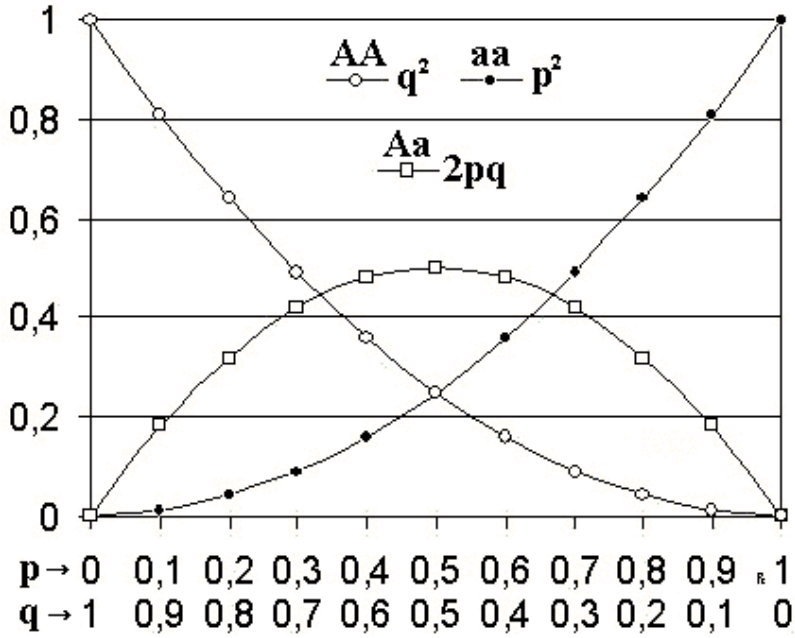


Imagen 21 El principio de Hardy-Weinberg para dos alelos: el eje horizontal muestra las dos frecuencias alélicas p y q, el eje vertical muestra la frecuencia de los genotipos y los tres posibles genotipos se representan por los distintos glifos.

En 1908, G.H. Hardy y W. Weinberg demostraron, por separado, que "en una población panmíctica (es decir, donde los individuos se aparean al azar), de gran tamaño y donde todos los individuos son igualmente viables y fecundos, el proceso de la herencia, por sí mismo, no cambia las frecuencias alélicas ni genotípicas de un determinado locus". En esencia, el principio de Hardy-Weinberg enuncia que, en ausencia de fuerzas, la descripción del sistema no cambia en el tiempo una vez alcanzado el equilibrio, y que la consecución de éste puede llevar una o más generaciones, dependiendo de las restricciones físicas impuestas por la organización del genoma.

La relación general entre frecuencias alélicas y genotípicas puede describirse en términos algebraicos: si p es la frecuencia de A1 y q es la de A2 , se cumple que p+q=1 si no existen más que esos dos alelos. Las frecuencias genotípicas de equilibrio vienen dadas por: p2 (A1 A1), 2pq (A1 A2), q2 (A2 A2). Por ejemplo, si p=0.6 y q=0.4, las frecuencias genotípicas son: p2 = 0.36 (A1 A1), 2pq=0.48 (A1 A2), q2 =0.16 (A2 A2). Obsérvese que las frecuencias genotípicas resultan del desarrollo de (p+q) (p+q)= (p+q)2 = p2 + 2pq + q2 . Con valores cualesquiera de p y q y con apareamiento aleatorio, una generación es suficiente para alcanzar el equilibrio en las frecuencias alélicas y genotípicas.

En una población en equilibrio, la frecuencia de heterocigotos es relativamente más alta cuanto más raro sea el fenotipo recesivo. Por ejemplo, una de cada 20,000 personas son albinas, por lo que, suponiendo equilibrio para el carácter, q2 = 1/20,000= 0.00005, y q = 0.007, y p = 0.993. En ese caso, 2pq = 0.014. Puesto que 0.014/ 0.00005 = 290, en las poblaciones humanas hay 290 veces más heterocigotos que homocigotos recesivos para el albinismo. Esto es una muestra de la dificultad de eliminar de las poblaciones los caracteres deletéreos recesivos, ya que la mayoría se encuentran en estado heterocigótico inexpresado y contra ellos no puede actuar la selección.

1.5.10.Heredabilidad

Heredabilidad es la proporción de la variación de caracteres biológicos en una población1 atribuible a la variación genotípica entre individuos. La variación entre individuos se puede deber a factores genéticos y/o ambientales. Los análisis de heredabilidad estiman las contribuciones relativas de las diferencias en factores genéticos y no-genéticos a la varianza

fenotípica total en una población. El valor de la heredabilidad indica en qué grado un rasgo o enfermedad se debe a causas genéticas o ambientales.

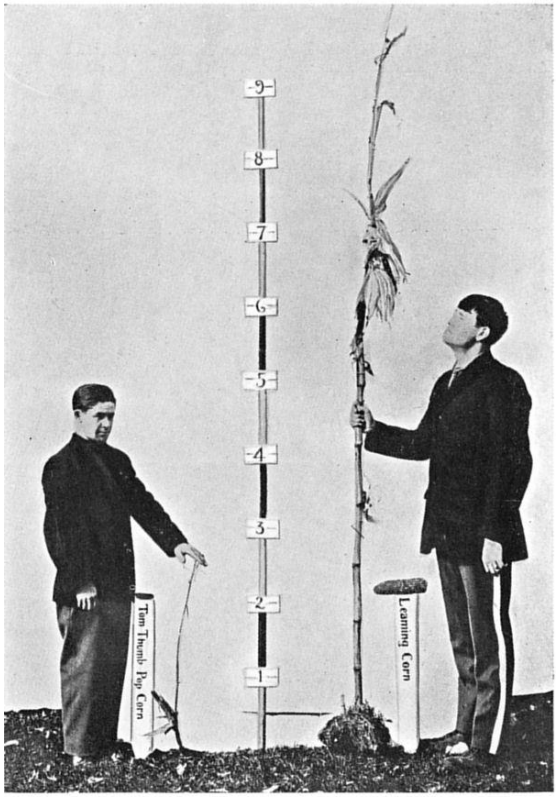
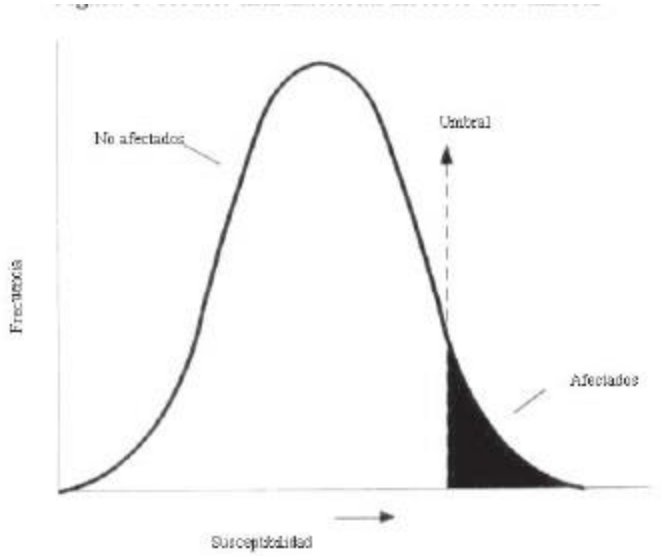


Imagen 22 Heredabilidad se hace preguntas como, por ejemplo, qué papel le toca a la genética en las diferencias en altura entre personas. Esto no es lo mismo que preguntar cómo la genética afecta la altura de un individuo.

Estos términos adicionales se pueden incluir en los modelos genéticos. Por ejemplo, el modelo genético más simple involucra un locus sencillo con dos alelos que afectan algún fenotipo cuantitativo, como lo muestra el ‘’+’’ de la Figura 1. Podemos calcular la regresión lineal del fenotipo sobre el número de alelos ‘’B’’ (0, 1, o 2), lo cual se muestra en la línea “Efecto Lineal”. Para cualquier genotipo, BiBj, el fenotipo esperado se puede escribir como la suma de la media total, el efecto lineal, y una desviación dominante Y cuantifica solo la porción de la variación fenotípica que es ‘’aditiva’’ (alélica) por naturaleza (nótese que para el sentido amplio se usa mayúscula H2, y para el sentido estricto, minúscula h2). Cuando el

interés es mejorar ganado vía selección artificial, por ejemplo, conocer la heredabilidad en sentido estricto del rasgo de interés, permite predecir qué tanto incrementará la media de la población en la próxima generación en función de qué tanto la media de los parentales seleccionados difiere de la media de la población de la cual fueron escogidos los parentales seleccionados. La ‘’respuesta a selección’’ observada conduce a una estimación de la heredabilidad en sentido estricto (llamada ‘’heredabilidad realizada’’).



Adaptado de Farrer y Cupples (17).

Imagen 23 La heredabilidad puede medirse en una curva de umbral la cual presenta dos fenotipos (el factor genético y el factor ambiental) y su expresión depende de una susceptibilidad llamado riesgo que varía continuamente.

Cuando la susceptibilidad es mayor que un valor umbral se expresa un rasgo especifico.

Las enfermedades se consideran características umbrales, si se presentan factores de susceptibilidad suficientes la enfermedad se desarrolla; de lo contrario no se presenta.

Geoffrey Miller, ha escrito sobre el tema de la selección sexual, «El concepto de la heredabilidad se aplica solo a esto caracteres que difieren entre individuos. Si un rasgo existiera en la misma forma entre toda una población, se puede llamar herencia pero no heredabilidad. (Wray & Visscher, 2015)

Uno de los mayores problemas de este tipo de enfermedades es cómo explicar su heredabilidad, ya que no es posible hacerlo por herencia mendeliana simple. Según el número de genes implicados, las enfermedades pueden clasificarse como monogénicas, oligogénicas y poligénicas, siendo el orden creciente. Es importante mencionar esto para poder entender el concepto de heredabilidad desaparecida. Algunos estudios recientes proponen que esta heredabilidad perdida puede ser menor a la que se creía en un principio, al producirse una interacción genética (epistasia) en los genes implicados.

Para las enfermedades provocadas por varios genes, una mutación en uno de ellos incrementará la probabilidad de desarrollarla, una mutación en otro gen volverá a incrementarla de manera aditiva y así sucesivamente hasta alcanzar un umbral (threshold en inglés) a partir del cual se considera que un individuo está afectado por la enfermedad. En este valor límite se deben tener en cuenta no sólo los factores genéticos, sino también los ambientales, de manera que cuando se supera un determinado número de factores acumulados, siempre se tendrá la enfermedad.

Así, se tiene que las enfermedades con un importante componente genético, aparte de poder estar influenciadas por factores externos, pueden estar relacionadas con mutaciones en otros loci no identificados que son, a su vez, los causantes de aquellas mutaciones

responsables en sí mismas del fenotipo alterado. Es aquí donde entra en juego el concepto de la heredabilidad desaparecida: la heredabilidad de algunos fenotipos sólo se explica genéticamente en un pequeño porcentaje.

Pero no sólo se puede hablar de heredabilidad desaparecida en cuanto a enfermedades, sino que hay otros ejemplos más comunes como puede ser la altura. Se sabe que la altura está determinada genéticamente, pero los estudios realizados sólo han conseguido explicarla en un 5% y relacionarla con 54 variantes diferentes. ¿Dónde está el resto del porcentaje de heredabilidad Es posible que existan condicionantes que hagan que se herede de forma no genética, como la dieta, el deporte, el nivel socioeconómico, etc. y que, además, haya más variantes implicadas aparte de las 54 que ya se conocen. Se puede decir que es más fácil predecir la altura que tendrá un bebé midiendo a los padres que teniendo en cuenta estudios genéticos.

CAPITULO 2. Planteamiento del problema

Aun cuando, en la actualidad, existe variedad de herramientas bioinformáticas desarrolladas para hacer análisis de genes y genomas, la cantidad de secuencias en las basase de datos publicas está incrementando de forma exponencial. La gran cantidad de información exige el desarrollo de nuevos métodos y algoritmos bioinformáticos, con capacidad de analizar grandes cantidades de datos. En el laboratorio de Bioinformática y Biofotónica del Instituto de Ingeniería de la UABC se han desarrollado, en los últimos años, distintos algoritmos y programas para realizar análisis de marcadores genéticos tipo SNP, sin embargo estos programas no tienen una interna gráfica y son ejecutados de forma independiente.

Debido a lo anterior en este proyecto de tesis de maestría se platea desarrollar una herramienta de software bioinformático, que integre los distintos algoritmos para hacer análisis de marcadores SNP, que contenga un ambiente gráfico y sea amigable tanto para usuarios expertos e inexperto, con el propósito de ser utilizados tanto en clases, como para realizar análisis para los distintos proyectos bioinformáticos del laboratorio.

2. Objetivos

2.1. Objetivo general.

Desarrollar una herramienta bioinformática, que contenga un ambiente gráfico amigable, e implemente algoritmos para realizar estimación de Frecuencias alélicas, Heterosigocidad y

Consanguinidad, partiendo de información de marcadores tipo SNP, de una población de individuos.

2.2. Objetivos específicos.

I. Obtención de un conjunto de genotipos tipo SNP de una población de individuos.

II. Implementación en código Python de los algoritmos para estimar frecuencias alélicas heterosigosidad y consanguinidad.

III. Desarrollo en código Python de una interfaz gráfica que integre los métodos implementados.

IV. Realizar pruebas de la herramienta, utilizando los datos reales de genotipos SNP.

CAPITULO 3. Herramienta de software para la estimación de parámetros genéticos.

3.1 Diseño de la herramienta.

Se pretende realizar una herramienta informática capaz de determinar el nivel de consanguinidad mediante la plataforma PYTHON de archivos de entrada .inp, la herramienta será capaz de cambiar el tipo de formato de entrada para realizar los cálculos correspondientes para determinar la consanguidad. La representación de los resultados actualmente es mediante consola. Se anexa ejemplo de interfaz gráfica para usuario con el fin de representación de datos a través de este interfaz.

3.1.1. Diagrama esquemático.

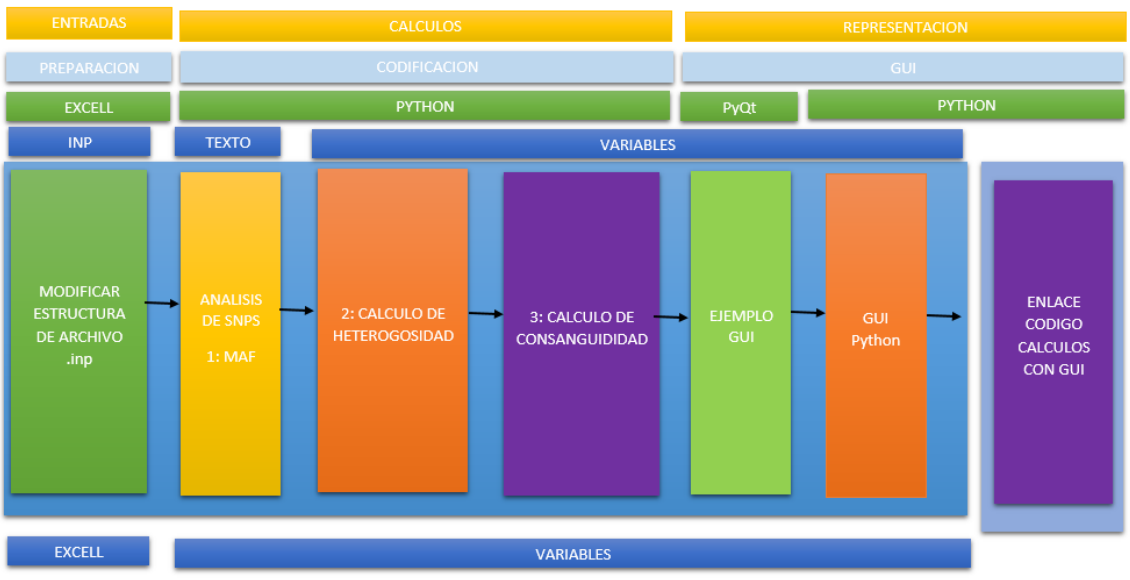


Imagen 24 Diagrama de bloques de Software

Descripción:

El diagrama anterior muestra el seguimiento que se tomó en cuenta para la realización del proyecto, donde la sección amarilla superior muestra el tipo de proceso que se realiza en esa sección, el segundo recuadro de la parte superior hasta la inferior, de color azul claro, determina el tipo de acción que se realizara en esta sección; la siguiente sección, color verde, determina el tipo de programa o lenguaje que se utiliza para la realización de dicho procedimiento en esta sección; la sección siguiente determina el tipo de documento de entrada en este proceso, color azul; y la siguiente sección determina el flujo de las funciones más específicas que se realizaran en el proceso correspondiente. Por último, se encuentra el tipo de variables de salida del proceso.

1: sección primaria.

a. Tipo de datos a manipular: ENTRADAS
b. Tipo de acción: PREPARACION
c. Tipo de lenguaje que se utilizó: EXCEL
d. Tipo de archivo a manejar: .inp
e. Acción: Modificación de estructura de archivo de entrada .inp
f. Tipo de documento de salida del proceso: Excell

2: sección secundaria.

a. Tipo de datos a manipular: CALCULOS
b. Tipo de acción: CODIFICACION
c. Tipo de lenguaje que se utilizó: PYTHON

d. Tipo de archivo a manejar: archivo de texto y variables de código PYTHON.

e. Acción 1: Análisis de datos y cálculo de MAF.

   Acción 2: Cálculo de Heterogosidad.

   Acción 3: Cálculo de Consanguinidad.

f. Tipo de documento de salida del proceso: Variables PYTHON.

3: sección terciaria.

a. Tipo de datos a manipular: REPRESENTACION

b. Tipo de acción: Interfaz gráfica.

c. Tipo de lenguaje que se utilizó 1: PyQt

   Tipo de lenguaje que se utilizó 2: PYTHON

d. Tipo de archivo a manejar: Variables

e. Acción 1: Creación de interfaz gráfica de ejemplo.

   Acción 2: Crear archivo .py de interfaz.

   Acción 3: Enlazar código de proceso con interfaz .py

f. Tipo de documento de salida del proceso: Gráficos.

3.1.2. Lenguaje de programación y código.

Python es un lenguaje de programación poderoso y fácil de aprender. Cuenta con estructuras de datos eficientes y de alto nivel y un enfoque simple pero efectivo a la programación orientada a objetos. La elegante sintaxis de Python y su tipificado dinámico, junto con su naturaleza interpretada, hacen de éste un lenguaje ideal para “scripting” y

desarrollo rápido de aplicaciones en diversas áreas y sobre la mayoría de las plataformas. El intérprete de Python y la extensa biblioteca estándar están a libre disposición en forma binaria y de código fuente para las principales plataformas desde el sitio web de Python, http: //www.python.org/, y puede distribuirse libremente. El mismo sitio contiene también distribuciones y enlaces de muchos módulos libres de Python de terceros, programas y herramientas, y documentación adicional. El intérprete de Python puede extenderse fácilmente con nuevas funcionalidades y tipos de datos implementados en C o C++ (u otros lenguajes accesibles desde C). Python también puede usarse como un lenguaje de extensiones para aplicaciones personalizables. (Rossum, 2009)

Código.

Sección primaria. Preparación de datos de entrada de documento de entrada en extensión .inp

```
Cells.Replace What:="~?", Replacement:="9" ' VALOR 9 PARA EVITAR ERROR
For i = 1 To 5
    Cells(i).EntireRow.Delete
    Next i
Sub separarDigitos()
For Each cell In Selection.Cells
cell.Select
    For i = 1 To Len(ActiveCell.Value)
        ActiveCell.Offset(0, i).Value = Mid(ActiveCell.Value, i, 1)
    Next i
Next cell
Cells.Replace What:="9", Replacement:="-9" 'CORREGIR VALOR A -9
End Sub
```

Sección secundaria. Cálculo MAF.

```
    # CALCULAR MAF : ALELO MENOR SOBRE ALELO MAYOR
    if conta != 0:
        if contc!=0:
            ame = min(conta,contc)
            if ame == conta:
                ALME.append("A")
            else:
                ALME.append("C")
            MAF= min(conta,contc)/max(conta,contc)
            VMAF.append(MAF)
        elif contg!=0:
            ame = min(conta,contg)
            if ame == conta:
                ALME.append("A")
            else:
                ALME.append("G")
```

```
        MAF= min(conta,contg)/max(conta,contg)
        VMAF.append(MAF)
    elif contt!=0:
        ame = min(conta,contt)
        if ame == conta:
            ALME.append("A")
        else:
            ALME.append("T")
        MAF= min(conta,contt)/max(conta,contt)
        VMAF.append(MAF)
    else:
        ame = 0
        ALME.append("X")
        MAF = 0
        VMAF.append(MAF)
elif contc != 0:
    if contg!=0:
        ame = min(contc,contg)
        if ame == contc:
            ALME.append("C")
        else:
            ALME.append("G")
        MAF= min(contc,contg)/max(contc,contg)
        VMAF.append(MAF)
    elif contt!=0:
        ame = min(contc,contt)
        if ame == contc:
            ALME.append("C")
        else:
            ALME.append("T")
        MAF= min(contc,contt)/max(contc,contt)
        VMAF.append(MAF)
    else:
        ame = 0
        ALME.append("X")
        MAF = 0
        VMAF.append(MAF)
elif contg != 0:
    if contt!=0:
        ame = min(contg,contt)
        if ame == contg:
            ALME.append("G")
        else:
            ALME.append("T")
        MAF= min(contg,contt)/max(contg,contt)
        VMAF.append(MAF)
    else:
        ame = 0
        #print("AMenor:",ame)
        ALME.append("X")
        MAF = 0
        VMAF.append(MAF)
else:
    ame = 0
    ALME.append("X")
    MAF = 0
    VMAF.append(MAF)
almy = max(conta,contc,contt,contg)
if almy == conta:
    ALMY.append("A")
    almy = 0
elif almy == contc:
    ALMY.append("C")
    almy = 0
```

```
    elif almy == contt:
        ALMY.append("T")
        almy = 0
    else:
        ALMY.append("G")
        almy = 0
archivo = open("Vector_MAF1.csv","w")
archivo.write('%s\n'%snps)
archivo.write('%s\n'%VMAF)
archivo.write('%s\n'%ALMY)
archivo.write('%s\n'%ALME)
archivo.close()
#### PARA MANEJAR ARCHIVO EN PYTHON
### GUARDAR COMO TXT POR TABULACIONES
```

Sección terciaria. Calculo heterogosidad y calculo consanguinidad.

```
for x in range (0,53):
    lista=linea[z].split("\t")
    lista1=linea[zz].split("\t")
    for i in range (1,4):
        if lista[i]!= lista1[i]:
            conth = conth + 1
        elif lista[i] == vamy[i].strip():
            contAA = contAA + 1
        else:
            contaa = contaa + 1
    z = z + 2
    zz = z + 1
print("CANTIDADES OBSERVADAS")
print("SNP:",i)
print("Homocigoto Mayor:",contAA) # n1
print("Heterocigotos:",conth) # n2
print("Homocigoto Menor:",contaa) # n3
N = (contAA+conth+contaa)
print("Total de Individuos:",N) # N
print("\n")
print("Frecuencias Genotipicas")
D = contAA/N
H = conth/N
R = contaa/N

print("D = ",D)
print("H = ",H)
print("R = ",R)
print("\n")
print("Frecuencias Observadas")
p = ((2*contAA) + conth)/(2*N)
q = ((2*contaa) + conth)/(2*N)
print("p = ",p)
print("q = ",q)
print("\n")
print("CANTIDADES ESPERADAS")
EAA = N*(p**2)
EAa = 2*N*p*q
Eaa = N*(q**2)
print("EAA = ",EAA)
print("EAa = ",EAa)
print("Eaa = ",Eaa)
print("\n")
print("Prueba X2")
X2_1 =((contAA - EAA)**2/EAA)
```

```
X2_2 =((conth - EAa)**2/EAa)
X2_3 =((contaa - Eaa)**2/Eaa)
X2 = X2_1+X2_2+X2_3
print("X2 = ",X2)
print("\n")
print("HETEROSIGOSIDAD")
nucl = N * 533
Ho = N / nucl
He = 0.186302002
print("Ho = ",Ho)
print("\n")
print("CONSANGUINIDAD")
print("F = ",Ho+He)
```

Sección terciaria. GUI PYTHON.

```
from PyQt5 import QtCore, QtGui, QtWidgets
class Ui_MainWindow(object):
    def setupUi(self, MainWindow):
        MainWindow.setObjectName("MainWindow")
        MainWindow.resize(644, 526)
        MainWindow.setCursor(QtGui.QCursor(QtCore.Qt.IBeamCursor))
        MainWindow.setStyleSheet("QMainWindow {\n"
"    \n"
"    color: qradialgradient(spread:reflect, cx:1, cy:1, radius:0.5, fx:0.6468, fy:0.6468, stop:0.625
rgba(0, 0, 0, 255), stop:1 rgba(255, 171, 171, 255));\n"
"\n"
"}")
        self.centralwidget = QtWidgets.QWidget(MainWindow)
        self.centralwidget.setObjectName("centralwidget")
        self.tableView = QtWidgets.QTableView(self.centralwidget)
        self.tableView.setGeometry(QtCore.QRect(160, 19, 191, 361))
        self.tableView.setObjectName("tableView")
        self.scrollArea = QtWidgets.QScrollArea(self.centralwidget)
        self.scrollArea.setGeometry(QtCore.QRect(500, 40, 120, 80))
        self.scrollArea.setWidgetResizable(True)
        self.scrollArea.setObjectName("scrollArea")
        self.scrollAreaWidgetContents = QtWidgets.QWidget()
        self.scrollAreaWidgetContents.setGeometry(QtCore.QRect(0, 0, 118, 78))
        self.scrollAreaWidgetContents.setObjectName("scrollAreaWidgetContents")
        self.scrollArea.setWidget(self.scrollAreaWidgetContents)
        self.tableView_2 = QtWidgets.QTableView(self.centralwidget)
        self.tableView_2.setGeometry(QtCore.QRect(360, 220, 256, 192))
        self.tableView_2.setObjectName("tableView_2")
        self.listView = QtWidgets.QListView(self.centralwidget)
        self.listView.setGeometry(QtCore.QRect(360, 41, 131, 171))
        self.listView.setObjectName("listView")
        self.treeView = QtWidgets.QTreeView(self.centralwidget)
        self.treeView.setGeometry(QtCore.QRect(500, 120, 121, 81))
        self.treeView.setObjectName("treeView")
        self.spinBox = QtWidgets.QSpinBox(self.centralwidget)
        self.spinBox.setGeometry(QtCore.QRect(310, 390, 42, 22))
        self.spinBox.setObjectName("spinBox")
        self.horizontalSlider = QtWidgets.QSlider(self.centralwidget)
        self.horizontalSlider.setGeometry(QtCore.QRect(360, 420, 251, 22))
        self.horizontalSlider.setOrientation(QtCore.Qt.Horizontal)
        self.horizontalSlider.setObjectName("horizontalSlider")
        self.buttonBox = QtWidgets.QDialogButtonBox(self.centralwidget)
        self.buttonBox.setGeometry(QtCore.QRect(150, 390, 156, 23))

self.buttonBox.setStandardButtons(QtWidgets.QDialogButtonBox.Cancel|QtWidgets.QDialogButtonBox.Ok)
        self.buttonBox.setObjectName("buttonBox")
        self.checkBox = QtWidgets.QCheckBox(self.centralwidget)
```

```
        self.checkBox.setGeometry(QtCore.QRect(10, 229, 141, 16))
        self.checkBox.setObjectName("checkBox")
        self.checkBox_3 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_3.setGeometry(QtCore.QRect(10, 249, 141, 16))
        self.checkBox_3.setObjectName("checkBox_3")
        self.checkBox_4 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_4.setGeometry(QtCore.QRect(10, 269, 141, 16))
        self.checkBox_4.setObjectName("checkBox_4")
        self.checkBox_5 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_5.setGeometry(QtCore.QRect(10, 289, 141, 16))
        self.checkBox_5.setObjectName("checkBox_5")
        self.checkBox_6 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_6.setGeometry(QtCore.QRect(10, 310, 141, 16))
        self.checkBox_6.setObjectName("checkBox_6")
        self.checkBox_7 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_7.setGeometry(QtCore.QRect(10, 330, 141, 16))
        self.checkBox_7.setObjectName("checkBox_7")
        self.checkBox_8 = QtWidgets.QCheckBox(self.centralwidget)
        self.checkBox_8.setGeometry(QtCore.QRect(10, 350, 141, 16))
        self.checkBox_8.setObjectName("checkBox_8")
        self.pushButton_7 = QtWidgets.QPushButton(self.centralwidget)
        self.pushButton_7.setGeometry(QtCore.QRect(530, 440, 102, 43))
        self.pushButton_7.setStyleSheet("QPushButton{\n"
"border-bottom-color: rgb(255, 170, 255);\n"
"\n"
"\n"
"}")
        self.pushButton_7.setObjectName("pushButton_7")
        self.splitter = QtWidgets.QSplitter(self.centralwidget)
        self.splitter.setGeometry(QtCore.QRect(10, 19, 102, 138))
        self.splitter.setOrientation(QtCore.Qt.Vertical)
        self.splitter.setObjectName("splitter")
        self.pushButton = QtWidgets.QPushButton(self.splitter)
        self.pushButton.setStyleSheet("QPushButton{\n"
"border-bottom-color: rgb(255, 170, 255);\n"
"\n"
"\n"
"}")
        self.pushButton.setObjectName("pushButton")
        self.pushButton_5 = QtWidgets.QPushButton(self.splitter)
        self.pushButton_5.setStyleSheet("QPushButton{\n"
"border-bottom-color: rgb(255, 170, 255);\n"
"\n"
"\n"
"}")
        self.pushButton_5.setObjectName("pushButton_5")
        self.pushButton_6 = QtWidgets.QPushButton(self.splitter)
        self.pushButton_6.setStyleSheet("QPushButton{\n"
"border-bottom-color: rgb(255, 170, 255);\n"
"\n"
"\n"
"}")
        self.pushButton_6.setObjectName("pushButton_6")
        self.splitter.raise_()
        self.tableView.raise_()
        self.scrollArea.raise_()
        self.tableView_2.raise_()
        self.listView.raise_()
        self.treeView.raise_()
        self.spinBox.raise_()
        self.horizontalSlider.raise_()
        self.buttonBox.raise_()
        self.checkBox.raise_()
        self.checkBox_3.raise_()
```

```
self.checkBox_4.raise_()
self.checkBox_5.raise_()
self.checkBox_6.raise_()
self.checkBox_7.raise_()
self.checkBox_8.raise_()
self.pushButton_7.raise_()
MainWindow.setCentralWidget(self.centralwidget)
self.menubar = QtWidgets.QMenuBar(MainWindow)
self.menubar.setGeometry(QtCore.QRect(0, 0, 644, 21))
self.menubar.setObjectName("menubar")
self.menuGuardar = QtWidgets.QMenu(self.menubar)
self.menuGuardar.setObjectName("menuGuardar")
self.menuAbrir = QtWidgets.QMenu(self.menubar)
self.menuAbrir.setObjectName("menuAbrir")
self.menuNuevo = QtWidgets.QMenu(self.menuAbrir)
self.menuNuevo.setObjectName("menuNuevo")
self.menuEditar = QtWidgets.QMenu(self.menubar)
self.menuEditar.setObjectName("menuEditar")
self.menuVer = QtWidgets.QMenu(self.menubar)
self.menuVer.setObjectName("menuVer")
self.menuConfiguraciones = QtWidgets.QMenu(self.menubar)
self.menuConfiguraciones.setObjectName("menuConfiguraciones")
self.menuAsociacion = QtWidgets.QMenu(self.menuConfiguraciones)
self.menuAsociacion.setObjectName("menuAsociacion")
self.menuAyuda = QtWidgets.QMenu(self.menubar)
self.menuAyuda.setObjectName("menuAyuda")
MainWindow.setMenuBar(self.menubar)
self.statusbar = QtWidgets.QStatusBar(MainWindow)
self.statusbar.setObjectName("statusbar")
MainWindow.setStatusBar(self.statusbar)
self.actionAbrir = QtWidgets.QAction(MainWindow)
self.actionAbrir.setObjectName("actionAbrir")
self.actionGuardar_como = QtWidgets.QAction(MainWindow)
self.actionGuardar_como.setObjectName("actionGuardar_como")
self.actionGuardar_como_2 = QtWidgets.QAction(MainWindow)
self.actionGuardar_como_2.setObjectName("actionGuardar_como_2")
self.actionGuardar_todo = QtWidgets.QAction(MainWindow)
self.actionGuardar_todo.setObjectName("actionGuardar_todo")
self.actionImprimir = QtWidgets.QAction(MainWindow)
self.actionImprimir.setObjectName("actionImprimir")
self.actionGuardar_imagen = QtWidgets.QAction(MainWindow)
self.actionGuardar_imagen.setObjectName("actionGuardar_imagen")
self.actionCerrar = QtWidgets.QAction(MainWindow)
self.actionCerrar.setObjectName("actionCerrar")
self.actionSalir = QtWidgets.QAction(MainWindow)
self.actionSalir.setObjectName("actionSalir")
self.actionAtras = QtWidgets.QAction(MainWindow)
self.actionAtras.setObjectName("actionAtras")
self.actionEnfrente = QtWidgets.QAction(MainWindow)
self.actionEnfrente.setObjectName("actionEnfrente")
self.actionCortas = QtWidgets.QAction(MainWindow)
self.actionCortas.setObjectName("actionCortas")
self.actionPegar = QtWidgets.QAction(MainWindow)
self.actionPegar.setObjectName("actionPegar")
self.actionEliminar = QtWidgets.QAction(MainWindow)
self.actionEliminar.setObjectName("actionEliminar")
self.actionSeleccionar = QtWidgets.QAction(MainWindow)
self.actionSeleccionar.setObjectName("actionSeleccionar")
self.actionTabla = QtWidgets.QAction(MainWindow)
self.actionTabla.setObjectName("actionTabla")
self.actionHerramientas = QtWidgets.QAction(MainWindow)
self.actionHerramientas.setObjectName("actionHerramientas")
self.actionBotones = QtWidgets.QAction(MainWindow)
self.actionBotones.setObjectName("actionBotones")
```

```
self.consola = QtWidgets.QAction(MainWindow)
self.consola.setObjectName("consola")
self.actionPreferencias = QtWidgets.QAction(MainWindow)
self.actionPreferencias.setObjectName("actionPreferencias")
self.actionLinkage = QtWidgets.QAction(MainWindow)
self.actionLinkage.setObjectName("actionLinkage")
self.actionHaps = QtWidgets.QAction(MainWindow)
self.actionHaps.setObjectName("actionHaps")
self.actionHapMap = QtWidgets.QAction(MainWindow)
self.actionHapMap.setObjectName("actionHapMap")
self.actionPHASE = QtWidgets.QAction(MainWindow)
self.actionPHASE.setObjectName("actionPHASE")
self.actionPLINK = QtWidgets.QAction(MainWindow)
self.actionPLINK.setObjectName("actionPLINK")
self.actionHaplotipos = QtWidgets.QAction(MainWindow)
self.actionHaplotipos.setObjectName("actionHaplotipos")
self.actionMarcadores = QtWidgets.QAction(MainWindow)
self.actionMarcadores.setObjectName("actionMarcadores")
self.actionConsanginidad = QtWidgets.QAction(MainWindow)
self.actionConsanginidad.setObjectName("actionConsanginidad")
self.actionHeterogocidad = QtWidgets.QAction(MainWindow)
self.actionHeterogocidad.setObjectName("actionHeterogocidad")
self.actionDe_un_marcador = QtWidgets.QAction(MainWindow)
self.actionDe_un_marcador.setObjectName("actionDe_un_marcador")
self.actionHaplotipos_2 = QtWidgets.QAction(MainWindow)
self.actionHaplotipos_2.setObjectName("actionHaplotipos_2")
self.actionTest_de_permutacion = QtWidgets.QAction(MainWindow)
self.actionTest_de_permutacion.setObjectName("actionTest_de_permutacion")
self.menuNuevo.addAction(self.actionLinkage)
self.menuNuevo.addSeparator()
self.menuNuevo.addAction(self.actionHaps)
self.menuNuevo.addSeparator()
self.menuNuevo.addAction(self.actionHapMap)
self.menuNuevo.addSeparator()
self.menuNuevo.addAction(self.actionPHASE)
self.menuNuevo.addSeparator()
self.menuNuevo.addAction(self.actionPLINK)
self.menuAbrir.addAction(self.menuNuevo.menuAction())
self.menuAbrir.addAction(self.actionAbrir)
self.menuAbrir.addAction(self.actionGuardar_como_2)
self.menuAbrir.addSeparator()
self.menuAbrir.addAction(self.actionGuardar_todo)
self.menuAbrir.addAction(self.actionImprimir)
self.menuAbrir.addAction(self.actionGuardar_imagen)
self.menuAbrir.addSeparator()
self.menuAbrir.addAction(self.actionCerrar)
self.menuAbrir.addAction(self.actionSalir)
self.menuEditar.addAction(self.actionAtras)
self.menuEditar.addAction(self.actionEnfrente)
self.menuEditar.addSeparator()
self.menuEditar.addAction(self.actionCortas)
self.menuEditar.addAction(self.actionPegar)
self.menuEditar.addAction(self.actionEliminar)
self.menuEditar.addAction(self.actionSeleccionar)
self.menuVer.addAction(self.actionTabla)
self.menuVer.addAction(self.actionHerramientas)
self.menuVer.addAction(self.actionBotones)
self.menuVer.addAction(self.consola)
self.menuAsociacion.addAction(self.actionDe_un_marcador)
self.menuAsociacion.addAction(self.actionHaplotipos_2)
self.menuAsociacion.addAction(self.actionTest_de_permutacion)
self.menuConfiguraciones.addAction(self.actionPreferencias)
self.menuConfiguraciones.addAction(self.actionHaplotipos)
self.menuConfiguraciones.addAction(self.actionMarcadores)
```

```
        self.menuConfiguraciones.addAction(self.menuAsociacion.menuAction())
        self.menuConfiguraciones.addAction(self.actionConsanginidad)
        self.menuConfiguraciones.addAction(self.actionHeterogocidad)
        self.menubar.addAction(self.menuAbrir.menuAction())
        self.menubar.addAction(self.menuConfiguraciones.menuAction())
        self.menubar.addAction(self.menuEditar.menuAction())
        self.menubar.addAction(self.menuVer.menuAction())
        self.menubar.addAction(self.menuAyuda.menuAction())
        self.menubar.addAction(self.menuGuardar.menuAction())

        self.retranslateUi(MainWindow)
        QtCore.QMetaObject.connectSlotsByName(MainWindow)

    def retranslateUi(self, MainWindow):
        _translate = QtCore.QCoreApplication.translate
        MainWindow.setWindowTitle(_translate("MainWindow", "Analisis gentico"))
        self.checkBox.setText(_translate("MainWindow", "Entradas"))
        self.checkBox_3.setText(_translate("MainWindow", "MAF"))
        self.checkBox_4.setText(_translate("MainWindow", "Cantidades observadas"))
        self.checkBox_5.setText(_translate("MainWindow", "Frecuencias genotipicas"))
        self.checkBox_6.setText(_translate("MainWindow", "Frecuencias observadas"))
        self.checkBox_7.setText(_translate("MainWindow", "Heterogocidad"))
        self.checkBox_8.setText(_translate("MainWindow", "Consanginidad"))
        self.pushButton_7.setText(_translate("MainWindow", "SALIR"))
        self.pushButton.setText(_translate("MainWindow", "CARGAR HATO"))
        self.pushButton_5.setText(_translate("MainWindow", "Vector MAF"))
        self.pushButton_6.setText(_translate("MainWindow", "ANALISIS"))
        self.menuGuardar.setTitle(_translate("MainWindow", "Guardar"))
        self.menuAbrir.setTitle(_translate("MainWindow", "Documento"))
        self.menuNuevo.setTitle(_translate("MainWindow", "Nuevo"))
        self.menuEditar.setTitle(_translate("MainWindow", "Editar"))
        self.menuVer.setTitle(_translate("MainWindow", "Ver"))
        self.menuConfiguraciones.setTitle(_translate("MainWindow", "Analisis"))
        self.menuAsociacion.setTitle(_translate("MainWindow", "Asociacion"))
        self.menuAyuda.setTitle(_translate("MainWindow", "Ayuda"))
        self.actionAbrir.setText(_translate("MainWindow", "Abrir"))
        self.actionGuardar_como.setText(_translate("MainWindow", "Guardar"))
        self.actionGuardar_como_2.setText(_translate("MainWindow", "Documentos recientes"))
        self.actionGuardar_todo.setText(_translate("MainWindow", "Guardar"))
        self.actionImprimir.setText(_translate("MainWindow", "Imprimir"))
        self.actionGuardar_imagen.setText(_translate("MainWindow", "Guardar imagen"))
        self.actionCerrar.setText(_translate("MainWindow", "Cerrar"))
        self.actionSalir.setText(_translate("MainWindow", "Salir"))
        self.actionAtras.setText(_translate("MainWindow", "Atras"))
        self.actionEnfrente.setText(_translate("MainWindow", "Enfrente"))
        self.actionCortas.setText(_translate("MainWindow", "Cortar"))
        self.actionPegar.setText(_translate("MainWindow", "Pegar"))
        self.actionEliminar.setText(_translate("MainWindow", "Eliminar"))
        self.actionSeleccionar.setText(_translate("MainWindow", "Seleccionar"))
        self.actionTabla.setText(_translate("MainWindow", "LD"))
        self.actionHerramientas.setText(_translate("MainWindow", "Herramientas"))
        self.actionBotones.setText(_translate("MainWindow", "Marcadores"))
        self.consola.setText(_translate("MainWindow", "Tagger"))
        self.actionPreferencias.setText(_translate("MainWindow", "LD Plot"))
        self.actionLinkage.setText(_translate("MainWindow", "Linkage"))
        self.actionHaps.setText(_translate("MainWindow", "Haps"))
        self.actionHapMap.setText(_translate("MainWindow", "HapMap"))
        self.actionPHASE.setText(_translate("MainWindow", "PHASE"))
        self.actionPLINK.setText(_translate("MainWindow", "PLINK"))
        self.actionHaplotipos.setText(_translate("MainWindow", "Haplotipos"))
        self.actionMarcadores.setText(_translate("MainWindow", "Marcadores"))
        self.actionConsanginidad.setText(_translate("MainWindow", "Consanginidad"))
        self.actionHeterogocidad.setText(_translate("MainWindow", "Heterogocidad"))
        self.actionDe_un_marcador.setText(_translate("MainWindow", "De un marcador"))
```

```
        self.actionHaplotipos_2.setText(_translate("MainWindow", "Haplotipos"))
        self.actionTest_de_permutacion.setText(_translate("MainWindow", "Test de permutacion"))

if __name__ == "__main__":
    import sys
    app = QtWidgets.QApplication(sys.argv)
    MainWindow = QtWidgets.QMainWindow()
    ui = Ui_MainWindow()
    ui.setupUi(MainWindow)
    MainWindow.show()
    sys.exit(app.exec_())
```

3.2. Ejemplos del interfaz gráfico.

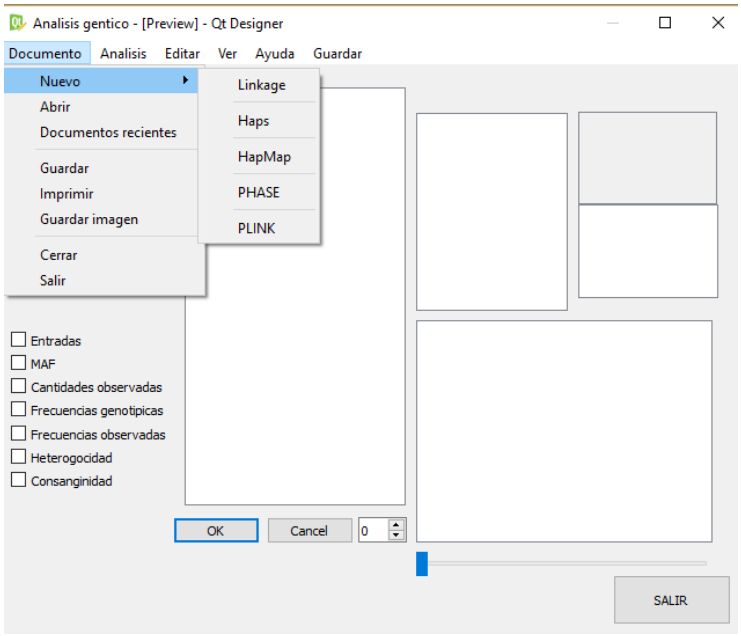
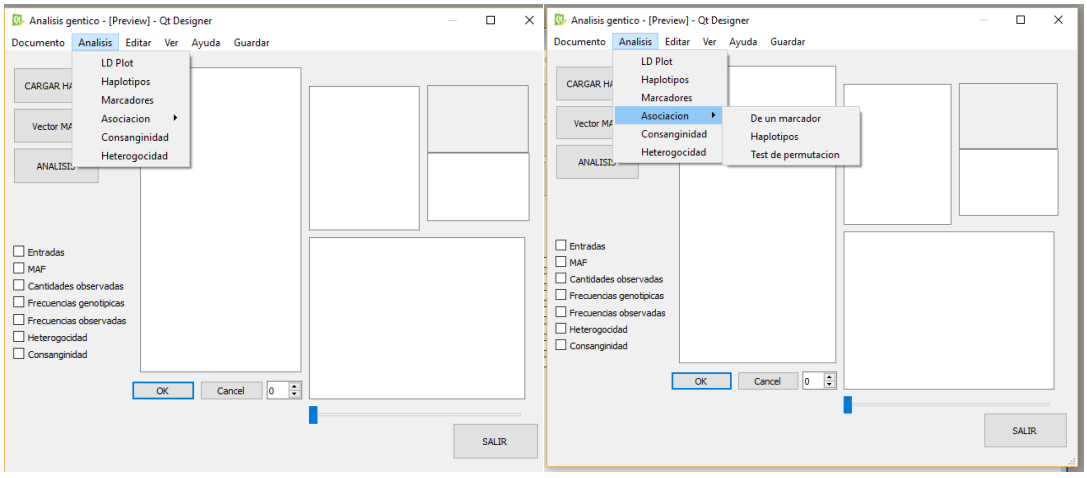


Imagen 25 Grafico principal de código .py de interfaz gráfica de sección terciaria.



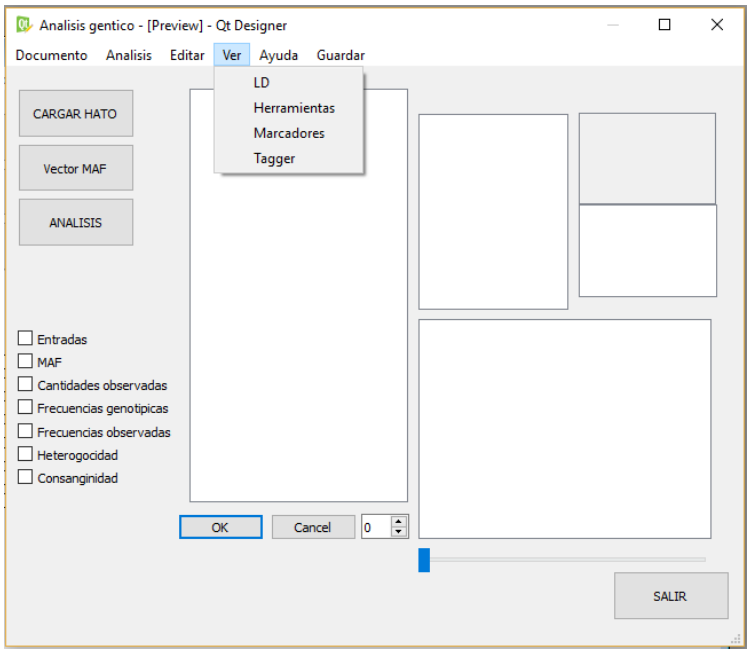


Imagen 26Graficos secundarios de interfaz gráfica de ejemplo.

3.3. Ejemplos de prueba.

Se utilizó el archivo de prueba HOL-29.inp, para realizar las pruebas de cálculo de MAF, Prueba X2, consanguinidad y heterosigocidad. Se codifico el proceso mediante un análisis en Excel del funcionamiento trabajando este documento HOL-29.

#HOL00048	A	C	C	T	C	C	T	C	T	-9	C
#HOL00049	A	C	C	C	C	A	G	T	T	C	C
#HOL00049	A	T	C	C	T	A	G	T	T	T	C
#HOL00050	A	T	T	T	C	C	T	C	T	C	C
#HOL00050	A	T	T	T	C	C	T	C	T	C	C
#HOL00051	A	C	C	C	T	A	G	T	T	T	C
#HOL00051	C	C	C	C	T	A	G	T	T	T	C
#HOL00052	A	C	T	T	C	C	T	C	T	C	C
#HOL00052	A	T	T	T	C	C	T	C	T	T	T
#HOL00053	A	C	T	T	C	A	G	C	T	-9	C
#HOL00053	A	C	T	T	C	C	T	T	T	-9	T
#HOL00055	A	C	T	C	C	A	G	C	T	C	C
#HOL00055	C	C	C	T	T	C	T	T	T	T	C
A	98	0	0	0	0	28	0	0	0	0	
C	8	78	51	50	84	78	0	78	0	50	
T	0	28	53	56	22	0	78	28	106	50	
G	0	0	0	0	0	0	28	0	0	0	
SNP	1	2	3	4	5	6	7	8	9	10	
MAF	0.08163265	0.35897436	0.96226415	0.89285714	0.26190476	0.35897436	0.35897436	0.35897436	0	1	0.10869
		38									

Imagen 27 Modelado de función de software, cálculo de MAF.

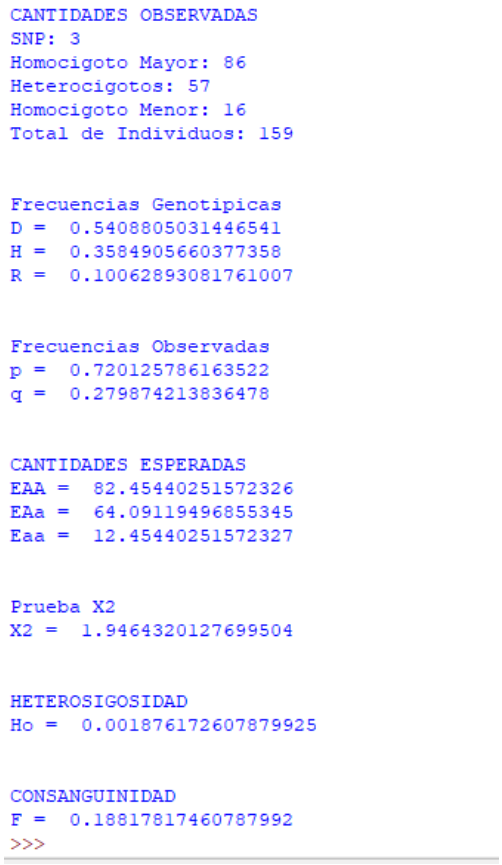


Imagen 28 Ejemplo de cálculos con código PYTHON. Aplicado a un solo SNP del archivo HOL-29. Se aplica a todos los SNP.

CAPITULO 4. Conclusiones y trabajo futuro.

4.1 Conclusiones.

En el planteamiento de este proyecto de tesis de maestría se emprendió como objetivo general desarrollar una herramienta bioinformática, que contuviera un ambiente gráfico amigable, e implementara algoritmos para realizar estimación de Frecuencias alélicas, Heterosigocidad y Consanguinidad, partiendo de información de marcadores tipo SNP, de una población de individuos. En el desarrollo de este documento, se presentan las etapas desarrolladas para el cumplimiento del objetivo general. Primero se presenta un estudio previo sobre conceptos básico de genética y genética de poblaciones, posteriormente se presenta el diseño de la herramienta bioinformática, incluyendo el código en el lenguaje de programación Python de los algoritmos para estimar frecuencias alélicas, heterosigosidad y consanguinidad. Igualmente se presenta un ejemplo de la interfaz gráfica desarrollada, y resultados obtenidos al probar la herramienta con datos de genotipos reales.

El trabajo presenta los primeros pasos para la generación de una herramienta bioinformática que englobe estimaciones de parámetros genéticos basada en marcadores SNP. En especial se consideraron la estimación de parámetros de Frecuencias alélicas, Heterosigocidad y consanguinidad.

4.2 Trabajo futuro.

Como trabajo futuro se plantea obtener una representacion gráfica de los datos, de una forma mas amigable para el usuario, y que integres el analisis de otros parametros genéticos, tales como Desequilibrio por ligamento, pruebas de asocacion, estmiacion de Valores Genomicos, entre otros.

Bibliografía

ACTEON. (s.f.). Obtenido de http://www.acteon.webs.upv.es/MATERIAL%20DOCENTE/Elementos%201.%20LD.pdf

B.J.B Keats, S. S. (2014). Population Genetics.

Barrett JC, F. B. (2005). *Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics.* PubMed ID: 15297300.

Britanica, E. (2018). *Marcador genetico.*

Camacho, J. C. (s.f.). Fundamentos de genetica de poblaciones.

Camacho., J. C. (2002). *Fundamentos de genetica de poblaciones.*

Can, T. (2014). *Methods in molecular biology (Clifton, N.J.).* Obtenido de NCBI: http://www.ncbi.nlm.nih.gov/pubmed/24272431

Clark, A. (2001). Population Genetics.

Desarrollo historico de la genetica humana. (s.f.).

DJ., S. J. (2016). *Statiscal Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguos. .*

Erson-Bensan, A. E. (2014). *miRNomics: MicroRNA Biology and Computational Analysis.* Springer New York Hedelberg Dordrecht London: Malik Yousef.

Excoffier, L. G. (2005). *Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics.*

Gallo, L. A. (2006). El uso de marcadores genéticos en el género Nothofagus con especial referencia a raulí y roble. *Bosque (Valdivia)*, 3-15.

Griffiths, A. J. (2005). *An Introduction to Genetic Analysis.*

Índice de fijación - Wikipedia, la enciclopedia libre. (s.f.). Obtenido de https://es.wikipedia.org/wiki/%C3%8Dndice_de_fijaci%C3%B3n

Jombart, T. (2008). *adegenet: a R package for the multivariate analysis of genetic markers, Bioinformatics.* Oxford University Press.

Martinez, J. M. (2004). Secuenciacion de genomas. *Arbor*.

nlm. (2 de Enero de 2019). *Genetics Home Reference*. Obtenido de Your Guide to Understanding Genetic Conditions: https://ghr.nlm.nih.gov/primer/genomicresearch/snp

Pinto, A. R. (s.f.). Introduccion a la genomica en VID. *GIE*.

Raquel, & GUINO, E. y. (2005). *Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos.* Obtenido de Scielo: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-91112005000400011&lng=es&nrm=iso

Rossum, G. v. (2009). *El tutorial de Python.* Fred L. Drake,Jr.

The Genomics Bottleneck. (2015). The Hutch Report.

Wray, N., & Visscher, P. (. (2015). Estimating Trait Heritability. *Nature Education* .

Xiong, J. (2006). *Essential Bioinformatics.* United States of America: Cambridge University Press, New York.