"Comparación de efectos epistáticos en interacción de polimorfismos (SNPs) y genes en enfermedades complejas"

María Gabriela Pizarro Inostroza

Máster en Bioinformática e Bioestadística
Bioestadística e Bioinformática (34)

Jaime Sastre Tomas
Carles Ventura Royo

02/01/2018

Ficha del Trabajo Final

Título del trabajo:	"Comparación de efectos epistáticos en interacción de polimorfismos (SNPs) y genes de enfermedades complejas"
Nombre del autor:	Mª Gabriela Pizarro Inostroza
Nombre del consultor/a:	Jaime Sastre Tomas
Nombre del PRA:	Carles Ventura Royo
Fecha de entrega (mm/aaaa):	02/01/2018
Titulación::	Máster universitario en bioinformática y bioestadística UOC -UB
Área del Trabajo Final:	Bioestadística e bioinformática
Idioma del trabajo:	Español
Palabras clave	Interacción; Epistasis; Enfermedades complejas

Resumen del Trabajo (máximo 250 palabras): Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.

La epistasis, comúnmente definida como la interacción entre genes, es un componente genético importante que subyace a la variación fenotípica y que pretende identificar las variables genéticas que están relacionadas o influyen en el riesgo de padecer una determinada enfermedad. En los últimos años se han desarrollado varios métodos estadísticos para modelar e identificar interacciones epistáticas entre variantes genéticas. Sin embargo, debido al gran espacio combinatorio de búsqueda de interacciones, a menudo la eficiencia del software utilizado no es suficiente para poder evaluar todas las posibles interacciones entre genes y SNPs. Por ello, en el presente estudio y con objeto de determinar interacciones entre genes o SNPs se compararon métodos estadísticos utilizando los software Plink y BEAM. Se estudiaron 1050 individuos, distribuidos en 525 casos y 525 controles con 500 SNPs. Los parámetros utilizados para comparar estos métodos son TRUE/FALSE POSITIVE/NEGATIVE, ya que los enfoques de regresión logística en combinación con un enfoque del modelo bayesiano facilitan un escenario más adecuado para el análisis e interpretación de datos en estudios de interacciones entre genes y SNPs. Los resultados obtenidos para los análisis de epistasis muestran que no hay interacciones significativas entre SNPs al analizar los datos con Plink, en cambio sí

para BEAM en los SNPs rs541178226611 y rs831225245351 asociados con mayor frecuencia en enfermedades relacionadas con carcinomas. En conclusión, el software BEAM es más exhaustivo que Plink para la búsqueda de interacciones entre SNPs.

Abstract (in English, 250 words or less):

The epistasis, commonly defined as the interaction between genes, is an important genetic component that underlies phenotypic variation and aims to identify the genetic variables that are related or influenced with a particular risk disease. In recent years various statistical methods have been developed to model and identify genetic variants between epistatic interactions. However, due to the large combinatorial space to search interactions, the efficiency of the software used is often not enough to evaluate all possible interactions between genes and SNPs. Therefore, in the present study and to determine interactions between genes or SNPs, statistical methods using Plink and BEAM software were compared. 1050 individuals, distributed in 525 cases and 525 controls with 500 SNPs were studied. The parameters used to compare these methods are TRUE / FALSE POSITIVE / NEGATIVE, since approaches to logistic regression in combination with a Bayesian model approach to facilitate a more appropriate setting for the analysis and interpretation of data in studies of interactions between genes and SNPs. The results obtained for the analysis of epistasis show that there is no significant difference between SNPs interactions to analyze data using Plink. However for BEAM, a interaction was observed in the SNPs rs541178226611 and rs831225245351 associated with greater frequency in diseases carcinomas-related. In conclusion, BEAM software is more exhaustive than Plink to determine interactions between SNPs.

Indice

1. Introducción

En los últimos años la epistasis, ha sido esencialmente importante para comprender tanto la estructura como la función de las vías genéticas y la dinámica evolutiva de los sistemas genéticos complejos.

La epistasis se puede definir como el efecto fenotípico de un locus que depende de uno o más loci, combinado de una o más variantes que puede dar lugar a un determinado fenotipo. Se habla comúnmente de interacción entre variantes de una sola base (SNPs) o interacciones de genes.

No obstante, la definición de epistasis en biología y estadística no son exactamente coherentes entre ellas. Es por eso que muchos investigadores definen la epistasis de diferentes formas (Moore H 2005; Cordell, 2002). Por ello, lo primero y fundamental que se debe distinguir en los estudios de epistasis son los dos tipos de que existen: la epistasis biológica (también llamada epistasis funcional), que fue descrita por Bateson (1909). En su definición original, sólo implicaba un efecto de un alelo en un locus oculto por el efecto de otro alelo en un segundo locus. Esto puede ser visto como una ampliación del concepto de dominancia a nivel inter-loci. Una definición más reciente también permite que los efectos de variantes genéticas sean mejorados por los efectos de otras variantes genéticas (Siemiatycki y Thomas, 1981), existiendo un efecto epistático cuando el efecto de un alelo en una variante genética depende de la presencia o ausencia de otra variante genética.

Por otro lado, encontramos la epistasis estadística, descrita por primera vez por Fisher (1918) que se refiere a la salida de los efectos aditivos de las variantes genéticas en diferentes loci con respecto a su contribución global al fenotipo (Cordell, 2002;Wang *et al*., 2010).

Uno de los objetivos primordiales a la hora de trabajar en interacciones epistáticas con métodos computacionales, es la interpretación de las interacciones encontradas estadísticamente relevantes, para así acercarse a lo descrito en la definición de biología y aprender el mecanismo funcional subyacente, aunque este último punto es indudablemente el más difícil (Moore y Williams, 2005) y es uno de los más ignorados en este tipo de estudios.

Una de las razones más comunes para detectar epistasis es principalmente la falta de éxito en la asociación directa entre genotipo y fenotipo. Podría ser que varios efectos conjuntos de la epistasis, determinan en parte el estado de la enfermedad (Mackay y Moore, 2014).

Es así que la detección de las interacciones entre genes que influyen en el riesgo de enfermedad es a través de interacciones complejas con otros genes y/o factores ambientales, lo cual sigue siendo un desafío estadístico y computacional (Templeton, 2000; Moore y Williams, 2002).

Se cree que las interacciones gen-gen son una fuente potencial de variación genética inexplicada (Eichler *et al*., 2010; Gibson, 2010; Manolio *et al.*, 2009; Zuk *et al*., 2012), pero siguen siendo en gran parte inexploradas. Una prueba de epistasis, es una prueba de si estos términos de interacción gen-gen (G × G) son cero o no, y la falta de epistasis representa una clase especial de todas las posibles funciones de penetrancia multi-locus. Por supuesto, es una cuestión empírica si la epistasis desempeña un papel mayor o menor para cualquier rasgo dado en cualquier población en particular o submuestra definida.

Un obstáculo importante para el estudio de la epistasis es la falta de algoritmos ampliamente aceptados que sean lo suficientemente rápidos para manejar eficazmente el polimorfismo de un solo nucleótido (SNPs) y en última instancia mejorar la comprensión del papel de la epistasis en la regulación genética de rasgos complejos (Niel, 2015).

En la actualidad, la identificación de las interacciones entre SNPs en todo el genoma son un desafío computacional y metodológico (Cordell, 2002), y son usadas para tratar de comprender mejor las características de enfermedades complejas.

La identificación de las posibles interacciones entre SNPs en los estudios genómicos, es una tarea importante cuando se investigan factores genéticos que influyen en rasgos complejos comunes (Arkin *et al*., 2014).

El avance de tecnologías de alto rendimiento genera millones de SNPs que pueden ser ampliamente explorados para detectar interacciones entre SNPs o genes. Sin embargo, la búsqueda de interacciones entre SNPs es un desafío complicado y requiere de herramientas computacionalmente eficientes y estadísticamente eficaces (Changshuai *et al*., 2014).

Esto supone dos grandes desafíos para enfrentar la detección de las interacciones entre SNPs en todo el genoma. El primer desafío es la carga computacional intensiva (Wei *et al*., 2014; Steen, 2010), teniendo que evaluar más de diez mil millones de combinaciones, aunque sólo se tengan en cuenta todas las posibles interacciones SNP entre pares. El otro es el desafío estadístico para lograr los umbrales de significación derivados después de la corrección de Bonferroni (Wei *et al*., 2014).

En los últimos años se han desarrollados varios métodos para estudiar datos genéticos que se centran en las interacciones epistáticas que se han propuesto en varias revisiones (Niel 2015; Wei, 2014; Steen, 2012). Estos métodos estadísticos incluyen frecuencias, métodos estadísticos, bayesianos y computacionales (Ritchie *et al.*, 2001; Zhang y Xu, 2005; Zhao y Xiong, 2006; Ferreira *et al.*, 2007; Zhang y Liu 2007; Gayan *et al.* , 2008; Li *et al.*, 2008; Park y Hastie, 2008; Jung *et al.* , 2009; Miller *et al.* , 2009; Wang, 2009; Wu *et al.* ,2009). Sin embargo, casi todos estos métodos estadísticos se centran en la búsqueda explícita de pares o de orden superior para identificar los efectos de las interacciones epistáticas. Teniendo en cuenta el espacio de búsqueda extremadamente grande (por ejemplo, p (p - 1) / 2 combinaciones de pares para p variantes), estos métodos a menudo sufren de carga computacional y baja potencia estadística. A pesar de diversas implementaciones computacionales eficientes (Hemani *et al*., 2011, Wan *et al*., 2010) y el desarrollo de búsqueda de algoritmos eficientes (Prabhu *et al*., 2012), la exploración sobre un gran espacio de búsqueda combinatoria sigue siendo una tarea complicada para los estudios de epistasis de gran tamaño.

1.1. Contexto y justificación

1.1.1. Contexto

Los estudios de asociación del genoma completo (GWAS) se están utilizando como estándar en los estudios de análisis genéticos de las enfermedades humanas complejas comunes como la diabetes o la hipertensión. Dada la complejidad y robustez de las redes biológicas, es poco probable que tales enfermedades sean el resultado de única mutación, sino más bien un conjunto de factores que interactúan entre ellos. Es por ello que en los últimos años la epistasis, ha sido esencialmente importante para comprender tanto la estructura como la función de las vías genéticas y la dinámica evolutiva de los sistemas genéticos complejos. La epistasis se puede definir como el efecto fenotípico de un locus que depende de uno o más loci, combinado de una o más variantes que puede dar lugar a un determinado fenotipo. Se habla comúnmente de interacción entre variantes de una sola base (SNPs) o interacciones de genes, los cuales confieren susceptibilidad a la enfermedad. Hasta la fecha, miles de SNPs se han asociado con enfermedades y otros rasgos complejos. Normalmente, el análisis estadístico busca la asociación entre un fenotipo y un SNPs tomado individualmente mediante pruebas de locus único, entendiendo como locus, un lugar específico del cromosoma donde se localiza un gen. Sin embargo, los genetistas admiten que este es un enfoque simplista para abordar la

complejidad de los mecanismos biológicos. Sin embargo, la capacidad de identificar las interacciones epistáticas en la práctica enfrenta desafíos estadísticos y computacionales importantes.

Los métodos estadísticos estándar exploran a través de las interacciones, resultando lento tanto computacionalmente como estadísticamente por el gran número de combinaciones de las interacciones. En este estudio propondremos un enfoque computacional, validando algoritmos con análisis de datos simulados, considerando pares de SNPs y proporcionando evidencia sobre las propiedades de las interacciones epistáticas. Algunos de los programas que se pretenden estudiar son: Bayesian Epistasis Association Mapping (BEAM), que es un modelo bayesiano para detectar asociaciones de enfermedad de locus únicos e interacciones multilocus en estudios de casos y controles, que se calcula a través de la cadena de Markov Monte Carlo (MCMC). La razón fundamental detrás del modelo BEAM es que, si algunos SNPs están asociados con la enfermedad, la distribución de sus genotipos (o alelos) debe ser diferente entre los casos y los controles, de lo contrario no hay evidencia de asociación de la enfermedad en estos SNPs. Para distinguir entre asociaciones interactivas y marginales de múltiples SNPs el modelo BEAM define un conjunto de SNPs para que sean interactivos si la distribución conjunta de estos SNPs se ajusta mejor a los datos que el modelo de independencia (es decir, el producto de sus respectivas distribuciones marginales), teniendo en cuenta que la "interacción" está bien definida sólo para aquellos SNPs que son mutuamente independientes (no vinculados) *a priori*, como los SNPs ubicados muy separados en el genoma. El algoritmo BEAM asigna todos los SNPs a tres grupos no solapantes: el grupo 1 contiene SNPs que están marginalmente asociados con la enfermedad, el grupo 2 contiene SNPs conjuntamente asociados con la enfermedad y el grupo 0 contiene SNPs que no están relacionados con la enfermedad. Una partición correcta de SNPs en los tres grupos es de interés directo para un estudio de asociación (Zhang e Liu, 2007). Otro de los programas de interés es Plink, siendo un programa que puede analizar grandes conjuntos de datos que comprenden cientos de miles de marcadores genotipados para miles de individuos que pueden ser rápidamente manipulados y analizados en su totalidad. Es un conjunto de herramientas C / C ++ de código abierto, para los estudios de GWAS y estudios de población. Proporciona herramientas para que los pasos analíticos básicos sean computacionalmente eficientes, admitiendo algunos enfoques novedosos para los datos del genoma completo. Los cinco dominios principales de la función de Plink son:

administración de datos, estadísticas de resumen, estratificación de la población, análisis de asociación y estimación de identidad por descendencia.

La regresión logística ha sido ampliamente utilizada como un método paramétrico para la búsqueda exhaustiva de interacciones en el análisis de asociación para detectar epistasis. Los conjuntos de datos GWAS muy grandes se pueden analizar utilizando hardware bastante estándar, y no hay límites fijos en la cantidad de muestras o SNPs. Por ejemplo, una estación de trabajo Linux con memoria de acceso aleatorio de 2 GB y un procesador dual de 3.6 GHz puede manejar> 5,000 individuos genotipados para 500,000 SNPs. Este programa fue desarrollado por Shaun Purcell en el centro de investigación genética humana (CHGR), el hospital general de Massachusetts (MGH) y el instituto Broad de Harvard y MIT(Purcell *et al*. ,2007). Es por eso que los parámetros más eficaces para comparar estos métodos son TRUE/FALSE POSITIVE/NEGATIVE, ya que los enfoques de regresión logística en combinación con un enfoque del modelo bayesiano nos proporciona un escenario más adecuado para el análisis e interpretación de datos en estudios de interacciones entre genes y SNPs.

1.1.2. Justificación

Los motivos que nos incentivaron a realizar este estudio fueron la identificación estadística de las interacciones entre SNPs y genes, que nos ayudarán a explicar la etiología genética de muchas enfermedades, mejorando así la comprensión de los factores genéticos y ambientales, además de su diagnóstico, prevención y tratamiento. En la actualidad se están llevando a cabo investigaciones desde el punto de vista económico y técnico de estudios de GWAS con más de un millón de SNPs, distribuidos a través del genoma. Lamentablemente, la detección de las interacciones entre genes y SNPs es extremadamente difícil debido al gran número de posibles interacciones, la ambigüedad con respecto a la codificación de marcadores y la escala de interacción. Para muchos conjuntos de datos, no hay suficiente poder estadístico para evaluar todas las posibles interacciones entre genes y SNPs. Estos estudios pretenden identificar las variables genéticas que están relacionadas o influyen en el riesgo de padecer una determinada enfermedad. Por ello, compararemos los métodos estadísticos de cada software seleccionado que nos permitirá observar y analizar las propiedades generales y las diferencias técnicas estadísticas, así como comprobar sus habilidades para detectar variables genéticas relacionadas con enfermedades.

1.2. Objetivos del trabajo

1.2.1. Objetivos generales

Objetivo 1: Se espera poder estudiar y probar métodos estadísticos y computacionales existentes actualmente para determinar interacciones entre genes o SNPs.

Objetivo 2: Se espera evaluar los diferentes métodos estadísticos, analizando los datos de interacciones entre genes y SNPs entregados por los programas estadísticos ya preseleccionados para este estudio.

1.2.2. Objetivos específicos

Objetivo 1: Estudiar los atributos de entrada y salida de los datos (en esta caso un conjunto de SNPs). Resaltando la capacidad de detectar interacciones de datos con alta dimensionalidad.

Objetivo 2: Estudiar el grado de calidad de la información que entregan los programas estadísticos utilizados para este estudio, determinando la posible existencia de correlación fenotipo-genotipo.

Objetivo 3: Estudiar y comparar los datos obtenidos en esta fase de recopilación de información, relacionándolos con los problemas de evaluación en la precisión y potencia estadística en la aplicación de interacciones entre genes y SNPs.

1.3. Enfoque y método a seguir

● Planificar, organizar y distribuir adecuadamente el tiempo que se debe tener para analizar cada programa y actividad a corto, mediano y largo plazo para alcanzar los objetivos de este estudio, teniendo en cuenta la gran cantidad de datos que se utilizarán en esta ocasión.

● Anticipar los posibles obstáculos que se presentarán en el desarrollo de este estudio para poder alcanzar los objetivos y corregir desviaciones que puedan surgir en el transcurso de este estudio.

● Tener claros los plazos de entrega del plan de trabajo.

1.4. Planificación del trabajo

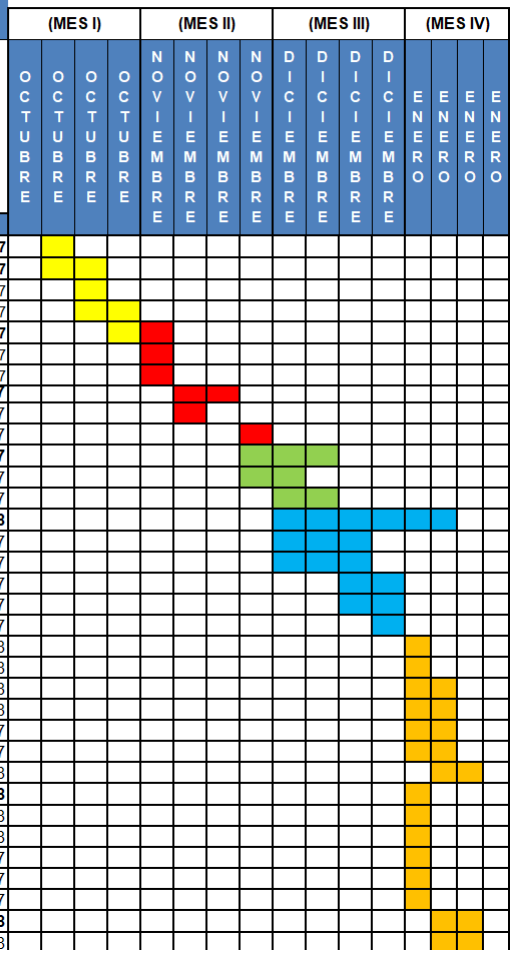
Para la correcta realización del TFM en el tiempo preestablecido para el semestre se ha organizado el trabajo en una serie de tareas temporalizadas, con el seguimiento por parte del profesor colaborador para garantizar la correcta ejecución del proyecto.

Plan de Trabajo

Proyecto:Trabajo Fin de Master
Fecha de inicio:20/09/2017
Fecha de término:02/01/2018

2017/2018

Tareas	fecha Inicio	Termino
Revision Bibliografica	**17/10/2017**	**22/10/2017**
Eleccion de Software	**18/10/2017**	**27/10/2017**
Plink caracteristicas	18/10/2017	22/10/2017
BEAM caracteristicas	23/10/2017	27/10/2017
funcionamiento software	**28/10/2017**	**04/11/2017**
Preparacion datos	05/11/2017	07/11/2017
formulas hojas para analisis	05/11/2017	07/11/2017
Analisis de datos	**08/11/2017**	**28/11/2017**
Analisis con Plink	08/11/2017	18/11/2017
Analisis con BEAM	20/11/2017	28/11/2017
Interpretacion de datos	**29/11/2017**	**18/12/2017**
Interpretacion con Plink	29/11/2017	12/12/2017
Interpretacion con BEAM	13/12/2017	18/12/2017
Desarrollo de la Memoria	**19/12/2017**	**21/01/2018**
Ficha Trabajo Final	19/12/2017	20/12/2017
Resumen	20/12/2017	21/12/2017
Introduccion	21/12/2017	24/12/2017
contexto y justificacion	26/12/2017	29/12/2017
objetivos del trabajo	30/12/2017	31/12/2017
Enfoque y metodologia	02/01/2018	06/01/2018
Planificacion	07/01/2018	09/01/2018
resultados	10/01/2017	14/01/2018
discusion	15/01/2018	17/01/2018
conclusiones	18/01/2017	19/01/2017
bibliografia	20/01/2018	21/01/2017
Anexos	20/01/2018	21/01/2018
Elaboaracion de la present	**03/01/2018**	**10/01/2018**
Diseño	03/01/2018	04/01/2018
introduccion	03/01/2018	04/01/2018
material y metod	05/01/2018	06/01/2017
resultados y discusión	05/01/2018	06/01/2017
conclusiones	07/01/2017	09/01/2017
defensa	**11/01/2018**	**22/01/2018**
Preparacion de la defensa	11/01/2018	22/01/2018



1.4.1. Seguimiento y control

El seguimiento de este TFM se ha realizado por el profesor colaborador de la UOC para la asignatura, quien se ha encargado de asegurar que el proyecto coincide con los objetivos propuestos. Este seguimiento se ha llevado a cabo a través de las entregas realizadas a lo largo del semestre, así como mediante la resolución de las dudas planteadas por el autor en momentos puntuales.

El sistema de control se ha basado en la elaboración de tres Pruebas de Evaluación Continua y una Entrega Final a la que han precedido varias entregas previas que han permitido realizar los ajustes finales siguiendo las instrucciones del profesor colaborador.

2. Sumario de productos obtenidos

2.1. Plink

Este programa puede analizar grandes conjuntos de datos que comprenden cientos de miles de marcadores genotipados para miles de individuos que pueden ser rápidamente manipulados y analizados en su totalidad. A continuación se mostrarán los pasos seguidos para obtener las diferentes salidas que nos proporciona el programa Plink para estudios de epistasis de casos y controles.

Entrada de archivos con extensión .ped/.map

- ***Archivo con extensión (. ped):*** ID familiar, individuo o identificación de la muestra, padre, madre, sexo y fenotipo con sus respectivas codificaciones.
- ***Archivo con extensión (.map):*** Cromosoma, identificación del marcador, posición genética y posición física con sus respectivas codificaciones (Este archivo nos entrega información del conjunto de variantes antes mencionado).

plink --ped data.ped /plink --map data.map

2.1.1. Análisis de epistasis

Para las muestras basadas en la población con riesgo de enfermedad, se realizó una entrada de epistasis para caso-control y solo caso, que son las diferentes posibilidades que el programa entrega para estos estudios. Los estudios de caso-control representan una estrategia muestral, en la que de manera característica se selecciona a la población en estudio con base en la presencia (caso) o ausencia (control) de la enfermedad. En cambio, el estudio de casos solo se limita a describir e identificar un conjunto de casos que han aparecido en un intervalo de tiempo. Estas pruebas solo son aplicables para muestras basadas en la población, no así a la familia. Para obtener los resultados en la prueba de epistasis se utilizó el siguiente comando: **plink --file data –epistasis.** Este análisis de epistasis hace una prueba predeterminada dependiendo si el fenotipo es un rasgo cuantitativo o binario, usando regresión lineal o logística. Para la regresión lineal se utiliza la siguiente fórmula: $Y = \beta_0 + \beta_1 \, g\,A + \beta_2 \, g\,B + \beta_3 \, g\,A \, g\,B$. Para cada par de variantes (A, B), donde g A y g B son recuentos de alelos; β_3 son coeficientes de pruebas de significación. De manera similar, dado un fenotipo de caso-control, que este en nuestro estudio, el comando utilizado es el mismo **--epistasis** pero en esta ocasión usamos una

regresión logística con la siguiente formula: ln (P (Y = caso) / P (Y = control)) = $\beta_0 + \beta_1$ g A + β_2 g B + β_3 g A g B. Esta prueba nos resulta muy útil para los casos en los cuales queremos predecir la ausencia o presencia de una característica según los valores de un conjunto de predictores, por lo que se puede aplicar a un rango más amplio de situaciones de investigación. Los coeficientes de regresión logística además pueden ser utilizados para estimar los odds ratio de cada variable independiente del modelo, siendo así los odds ratio una estimación de la asociación de un determinado factor con una enfermedad, por lo que resulta necesario calcular una medida de variabilidad de esta estimación. Siendo el intervalo de confianza donde se encuentra el verdadero valor de odds ratio, lo cual nos permite obtener una buena estimación cuando un odds ratio se aproxima a 1, pero se hace menos estable cuando el odds ratio es mayor que 1.

El comando **--epistasis** nos muestra una gran cantidad de comparaciones de SNPs por SNPs. Teniendo en cuenta que la salida puede contener millones de líneas de resultados, una de las salidas que encontramos es **plink .epi.cc** (cc: caso/control), que nos entrega el valor predeterminado, el cual utilizaremos para este análisis, según lo indicado en el programa para realizar las comparaciones entre SNPs. Los odds ratio como comentamos más arriba, son una estimación de la asociación de un determinado factor con una enfermedad, lo cual observando los valores obtenidos en este estudio tras analizar 500 SNPs y 1050 individuos distribuidos en 525 casos y 525 controles, los valores de odds ratio fueron mayores que 1, lo cual se interpreta como que no existe interacciones entre los SNPs y los fenotipos asociados a estos, encontrando valores de odds entre 115.89 y 478.837, lo que quiere decir que existe una probabilidad del 0.11% y 0.4 7 % de que los SNPs 1:,:248650422:C:T:OR2T27 (relacionado con el cáncer), 12:,:753986:G:A:WNK1 (relacionado con el síndrome de Gordon), 1:,:15716271:T:C:PLEKHM2 (relacionado con la miocardiopatía dilatada recesiva y 16:,:70177296:T:C:CLEC18C (relacionado con el tumor lipomatoso atípico), localizados en los cromosomas 1, 12 y 16 respectivamente interactúen con el fenotipo (Siendo cualquier característica o rasgo observable de un organismo, como su morfología, desarrollo, propiedades bioquímicas, fisiología y comportamiento).

Otro de los puntos importantes a destacar es la columna de p valor que nos indica que SNP está fuertemente asociado con los fenotipos en nuestro conjunto de datos. El programa por defecto calcula un p valor con un rango de 1e-4. No obstante, se realizó un nuevo análisis con el comando –epi incluyendo un p valor de <0.0001, intentando dejar

los SNPs más significativos en el estudio de caso/control. Los p valores obtenidos en este análisis fueron los siguientes: de 5.50e-0.2 hasta 6.99e-07, observando que no existe una diferencia significativa en la comparación entre los diferentes SNPs estudiados para caso/control, concluyendo que no hay relación existente entre los SNPs estudiados para análisis de caso/control.

La entrada **plink .epi.cc.summary**. (cc: caso/control), muestra una idea muy aproximada del alcance de la epistasis y los SNPs. Analizando el desequilibrio de ligamiento (LD), siendo la propiedad de que algunos genes de la población genética no segreguen de forma independiente. Esta tendencia se incrementa cuando dos genes están muy próximos entre sí, con lo que desciende la probabilidad de recombinación. De modo que estando dos genes próximos en un mismo cromosoma, la probabilidad de heredar en bloque los alelos provenientes de los gametos paterno y materno aumenta. Los mayores valores de chi cuadrado obtenidos implican una mayor interacción entre los SNPs. Los resultados se representan en la tabla 5 mostrando los 10 primeros SNPs que interactúan con mayor fuerza.

Tabla 5: Localización e interacción entre SNPs (caso/control).

Chr	SNPs	BEST_CHISQ	Chr	SNPs
3	3:.:195778917:G:C:MUC4	25.96	6	6:.:31356227:C:G:HLA-B
6	6:.:31356227:C:G:HLA-B	25.96	3	3:.:195778917:G:C:MUC4
7	7:.:76515166:A:G:UPK3B	25.9	21	21:.:42111698:A:C:UMODL1
21	21:.:42111698:A:C:UMODL1	25.9	7	7:.:76515166:A:G:UPK3B
6	6:.:31356226:T:A:HLA-B	25.52	3	3:.:195778917:G:C:MUC4
1	1:.:248559206:C:T:OR2T29	25.42	15	15:.:40252416:G:A:C15orf56
15	15:.:40252416:G:A:C15orf56	25.42	1	1:.:248559206:C:T:OR2T29
3	3:.:13571436:A:G:FBLN2	25.39	19	19:.:36039441:C:T:THAP8
19	19:.:36039441:C:T:THAP8	25.39	3	3:.:13571436:A:G:FBLN2
1	1:.:15716271:T:C:PLEKHM2	25.19	4	4:.:8601390:A:T:CPZ

Chr: Cromosoma; SNPs: Polimorfismo de nucleótido único; BEST_CHISQ: chi cuadrado.

Ahora bien, los rasgos de enfermedad pueden usar un método aproximado pero más rápido para detectar epistasis, donde utilizamos el comando **--fast-epistasis**, aunque hay que tener en cuenta que solo se utiliza en análisis epistáticos de casos (población enferma). Para esto utilizamos el siguiente comando, **plink –file data --fast-epistasis --case-only.** Estos comandos nos entregan dos salidas de archivos con las siguientes extensiones: **plink.epi.co y plink.epi.co.summary**. (co: caso). Al igual que los estudios de caso/control, nos encontramos en la primera salida que no existe diferencia significativa en este estudio, encontrándonos p valores de 1,77e-06 hasta 2.22e-04, lo cual

implica que no existe una relación entre los SNPs para el estudio de solo caso. La segunda salida ya mencionada nos proporciona los mayores valores de chi cuadrado, lo cual nos indica una mayor interacción entre los SNPs. Los resultados obtenidos se mostrarán en la tabla 6, donde se muestran los 10 SNPs con mayor interacción.

Tabla 6: Localización e interpretación de SNPs (casos).

Chr	SNPs	BEST_CHISQ	Chr	SNPs
1	1:.:113940529:T:C:HIPK1	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:13176324:C:G:PRAMEF9	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:13308531:A:G:PRAMEF33P	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:146988091:A:G:NBPF12	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:152798045:G:A:LCE1D	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:152798057:G:A:LCE1D	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:15716271:T:C:PLEKHM2	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:158845615:T:G:MNDA	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:166989473:T:G:MAEL	36.21	2	2:.:113241577:A:G:PAX8
1	1:.:16758533:G:A:MST1L	36.21	2	2:.:113241577:A:G:PAX8

Chr: Cromosoma; SNPs: Polimorfismo de nucleótido único; BEST_CHISQ: chi cuadrado.

2.1.2. Análisis de asociación

La prueba de asociación de casos y controles compara genotipos entre dos grupos de fenotipos, un grupo llamado control que son los que no presentan la enfermedad y otro grupo llamado caso, en los cuales los sujetos si presentan la enfermedad. Estos grupos se comparan con la finalidad de conocer si estan o no asociadas con nuestros SNPs en estudios. Esta prueba de asociación realiza cinco modelos diferentes para comprobar la asociación de fenotipos binarios con genotipos bialélicos, siendo estas la prueba de asociación alélica, la prueba de asociación de tendencia Cochran-Armitage. Esta prueba está basada en el modelo de regresión lineal, la cual se trata de una prueba de dos muestras para las diferencias entre los casos y los controles con respecto al número promedio de alelos de riesgo que ocurren en el genotipo de un individuo. La prueba de asociación genotípica (prueba 2df), proporciona una prueba general de asociación en una tabla de 2 por 3 de enfermedad por genotipo. Los modelos dominantes y recesivos son pruebas para el alelo menor, es decir, D es el alelo menor y d es el alelo principal (tabla 7).

Tabla 7. Pruebas genotípicas.

Prueba	Alelos
Alélica	D/d
Dominante	DD, Dd /dd
Recesivo	DD/Dd, dd
Genotípico	DD/Dd/dd

Los comandos que se utilizaron para este apartado son los siguientes: **--assoc, --fisher y --model**. El comando de entrada es **plink --file data –assoc,** el cual nos entrega un archivo con extensión **plink.assoc.** Donde encontramos que la mayor frecuencia de alelos con respecto a los casos fue de 0.0009524. Por otro lado encontramos la mayor frecuencia alélica con respecto a los controles siendo este de 0.4057.

Se realizó la prueba exacta de Fisher, siendo utilizado para obtener los valores p para ver si existe significancia en el análisis de asociación de caso y control. Estos análisis se tratan de estudios analíticos para investigar causalidad, además se emplean con frecuencia para identificar factores de riesgo que se asocian causalmente con las enfermedades. Para la prueba de Fisher utilizamos el siguiente comando **plink --file data –Fisher.**, el cual crea un archivo de salida con la siguiente extensión: **plink.fisher**. Aquí utilizamos una significancia del 0.05 %, lo cual observando el p valor obtenido, los resultados muestran que para los estudios de caso/control son altamente significativos, como mostraremos en la tabla 8.

Tabla 8: SNP representativos y su localización.

CHR	SNP	p valor
14	14:.:104949318:G:C:AHNAK2	2,11E-149
7	7:.:100956998:C:A:MUC3A	5,46E-146
6	6:.:31356227:C:T:HLA-B	2,25E-131
20	20:.:62565060:T:C:MIR1-1HG	2,51E-129
8	8:.:3052470:G:A:CSMD1	1,65E-128
12	12:.:9984958:A:C:CLEC12A	1,78E-126
6	6:.:31356226:T:G:HLA-B	7,41E-125
13	13:.:39655754:G:A:COG6	1,24E-119
7	7:.:100956991:A:G:MUC3A	3,10E-119
9	9:.:128822790:A:G:C9orf114	3,10E-119

Por último se realizó una prueba de asociación entre una enfermedad, utilizando la opción –model. Para esto utilizamos el siguiente comando: **plink --file mydata –model,** el cual nos entrega un archivo de salida con extensión **plink.model**. Esta prueba nos facilita cinco pruebas diferentes, las cuales son: Allelic (alélica), Trend (Cochram –Armitage prueba de tendencia), Geno (genotipo), Dom (dominante) o Rec (recesiva). Los conteos genotípicos o alélicos se dan para casos y controles por separado. Para las pruebas recesivas y dominantes, los recuentos representan los genotipos, con dos de las clases agrupadas. Aquí encontramos que el test con mayor significancia con un p valor de 2.11e-149 tiene una frecuencia de alelos de casos versus controles de 0/1050 – 426/624. Ahora

bien, el test con mayor significancia fue de 9.74e-81 con una frecuencia entre casos versus controles de 0/1/524 – 32/196/297.

2.2. BEAM

Se utilizó el software BEAM. Es un enfoque bayesiano para detectar asociaciones de enfermedad de locus únicos e interacciones multilocus en estudios de caso y control que se calcula a través de la cadena de Markov Monte Carlo (MCMC), siendo un método de simulación para generar muestras de las distribuciones a posteriori y estimar cantidades de interés a posteriori. Bajo un enfoque bayesiano, no existe diferencia conceptual entre parámetros y valores observables, es decir, son en su totalidad cantidades aleatorias. Se realiza inferencia sobre una distribución a posteriori $\pi(\theta|x)$. La razón fundamental detrás del modelo BEAM es que, si algunos SNPs están asociados con la enfermedad, la distribución de sus genotipos (o alelos) deben ser diferentes entre los casos y los controles, de lo contrario no hay evidencia de asociación de la enfermedad en estos SNPs. En resumen este es un procedimiento de prueba de hipótesis que prueba cada marcador para detectar interacciones significativas. Utilizando la siguiente fórmula:

$$B_M = ln\frac{P_A(D_M, U_M)}{P_0(D_M, U_M)} = ln\frac{P_{join}(D_M)[P_{ind}(U_M) + P_{join}(U_M)]}{P_{ind}(D_M, U_M) + P_{join}(D_M, U_M)}$$

Donde M representa cada conjunto de k marcadores, que representan diferentes complejidades de las interacciones; D_M y U_M son datos de genotipo de M casos y controles; P_0 (D_M; U_M) y P_A (D_M; U_M) son los factores Bayesianos; P_{ind} es la distribución que supone independencia entre los marcadores en M; P_{join} es una distribución conjunta saturada de combinaciones de genotipos entre todos los marcadores en M.

Entrada del archivo con extensión txt.

Archivo txt: Contiene las siguientes columnas: Id SNPs, cromosoma, posición, caso y control.

Para ejecutar el programa se debe ingresar el siguiente comando para posicionarlo en la carpeta en la cual vamos a trabajar.

cd /home/osboxes/desktop/beam

Una vez ya ejecutado el programa se procedió a ingresar el archivo txt preparado con el siguiente comando.

./BEAM BEAMData.tx

Este comando nos entrega 3 archivos de salida con extensión .pro, .dot y .chi.

2.2.1. Archivo de salida pry.pro

Nos entrega las probabilidades posteriores de asociaciones marginales e interacción por SNPs. Además se pueden sumar las probabilidades de múltiples SNPs en una región para estimar el número de SNPs asociados a la enfermedad. Este archivo de salida enumera los SNPs del 0 al 499.

En los resultados se observa que en la mayoría de los datos no existe ninguna asociación marginal e interacción entre los SNPs, exceptuando en los SNPs rs541178226611, rs831225245351 y rs2361955083628, ubicados en los cromosomas 11, 12 y 19, encontrados en carcinomas (tabla 9), donde encontramos que existe una probabilidad posterior en tales posiciones de 0.495,1.00 y 0.005 respectivamente.

Tabla 9. Probabilidades posteriores de asociaciones marginales e interacciones por SNPs (pry.pro).

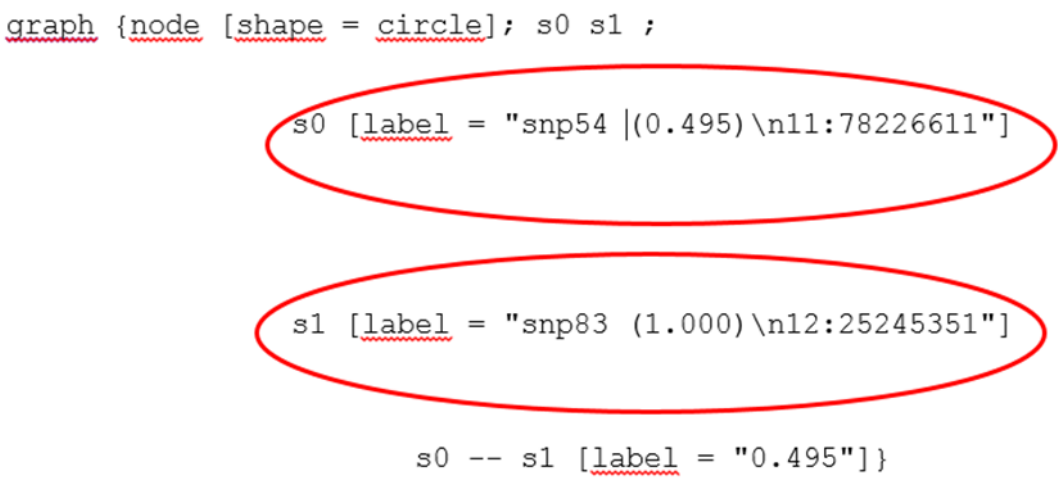
ID	Chr	Posición	Marginal		Interacción		Total	Posterior
rs54	11	78226611	0.000000	+	0.495000	=	0.495000	0.495000
rs83	12	25245351	0.500000	+	0.500000	=	1.000.000	1.000.000
rs236	19	55083628	0.005000	+	0.000000	=	0.005000	0.005000

Id: Identificación del individuo; Chr: Cromosoma; Posición: Posición en el cromosoma; Margina: Asociación marginal; interacción: Interacción por SNP.

2.2.2. Archivo de salida pry.dot

Nos entrega el gráfico de asociación primaria tomando en cuenta el desequilibrio de ligamiento desconocido entre los SNPs, entregándonos resultados sensibles y específicos para este estudio. Las primeras líneas en el archivo definen nodos gráficos, y los números en el paréntesis indican la probabilidad posterior de asociación del nodo (sumada a la región de 100 kb). Las últimas líneas en el archivo definen los borde, donde encontramos que los SNPs rs541178226611 y rs 831225245351 indican una probabilidad de asociación de la enfermedad de 0.495 y 1.000, encontrados en carcinomas, confirmándose que los dos SNPs están asociados a la enfermedad.

Gráfico 1. Asociación de SNPs a la enfermedad.



2.2.3 Archivo de salida chi.txt

Es un archivo de prueba de chisq single-SNPs, con estadísticos de prueba y conteo de alelos. Podemos decir que los alelos encontrados con mayor frecuencia tanto para caso como para controles se encuentran en la posición 122328146 y 104949318 dentro del cromosoma 12 y 14, encontrándose asociados a enfermedades del miocardio como observamos en las tablas 10 y 11. Por otro lado podemos observar que los alelos con menor frecuencia en estudios de caso y controles los encontramos en la posición 152213566 y 112862509 dentro de los cromosomas 1 y 5. Encontrados en asociaciones en enfermedades de la piel y carcinoma. Por último, en la tabla 12 y 13 encontramos alelos con menor frecuencia tanto en caso como para control, encontrando un valor que se repite en la posición 25245351 dentro del cromosoma 12, encontrados en asociaciones en tumores.

Tabla 10. Conteo de alelos con mayor frecuencia en casos (pry.chi).

ID	Chr	Posición	Chi	Alelo	Caso	Alelo	Caso
rs74	12	122328146	3.905.617	343	182	352	143
rs179	17	45983409	4.760.435	349	143	330	148

Id: Identificación del individuo; Chr: Cromosoma; Chi: Chi cuadrado; Alelo: Alelos con mayor frecuencia; Caso: Enfermedad.

Tabla 11. Conteo de alelos con mayor frecuencia en controles.

ID	Chr	Posición	Chi	Alelo	Control	Alelo	Control
rs116	14	104949318	0.647979	382	120	367	111
rs433	17	45983409	4.760.435	349	33	330	17

Id: Identificación del individuo; Chr: Cromosoma; Chi: Chi cuadrado; Alelo: Alelos con mayor frecuencia; Control: Ausencia de la enfermedad.

Tabla 12. Conteo de alelos con menor frecuencia en casos.

ID	Chr	Posición	Chi	Alelo	Caso	Alelo	Caso
rs30	1	152213566	0.137510	471	7	443	8
rs83	12	25245351	56.333.680	467	56	495	0

Id: Identificación del individuo; Chr: Cromosoma; Chi: Chi cuadrado; Alelo: Alelos con mayor frecuencia; Control: Ausencia de la enfermedad.

Tabla 13. Conteo de alelos con menor frecuencia en controles.

ID	Chr	Posición	Chi	Alelo	Control	Alelo	Control
rs83	12	25245351	56.333.680	467	2	495	0
rs358	5	112862509	0.340707	485	0	462	0

Id: Identificación del individuo; Chr: Cromosoma; Chi: Chi cuadrado

2.3. Comparación de los software utilizados para el estudio de interacción entre SNPs

El software PLINK (Purcell et al., 2007) ha implementado modelos de regresión logística para detectar epistasis. Pero, en datos de alta dimensión, la estimación de parámetros es un procedimiento costoso y no preciso que introduce grandes errores estándar porque los tamaños de muestra son demasiado pequeños en comparación con el tamaño de datos del genoma. Como consecuencia, se generan muchos falsos positivos cuando se trata de tales datos. Por lo tanto, la estrategia de regresión logística ha sido ampliamente descrita como inadecuada para manejar conjuntos de datos genómicos (Cordell, 2009; Moore y Williams, 2009; Steen, 2012). En nuestro estudio no se han hallado interacciones entre SNPs cuando se ha utilizado el software Plink. Esto puede ser consecuencia de que el programa no haya detectado interacciones porqué no existieron o porqué realmente el programa al trabajar con una muestra pequeña puede que su capacidad no sea suficiente para detectar las posibles interacciones que si se han observado cuando se ha utilizado el programa BEAM. El hecho de encontrar interacciones puede ser debido a que BEAM, al aplicar modelos bayesianos y permitir filtrar los marcadores significativos de acuerdo al umbral preestablecido de la probabilidad posterior simplifica el modelo, permitiendo con mayor precisión evaluar las asociaciones e interacciones entre SNPs.

3. Conclusiones

- Para los estudios estándar de caso/control, la estimación del desequilibrio de ligamiento dentro de los controles, proporciona el mismo resultado que en los casos, entonces el desequilibrio de ligamiento observado no se origina en las interacciones epistáticas. Su frontera con las interacciones epistáticas puede ser

borrosa ya que los programas de software están diseñados para detectar SNPs que afectan conjuntamente al fenotipo.

- No se destaca ninguna estrategia para detectar epistasis: Se deben buscar el uso de más programas para analizar los datos para lograr un equilibrio entre la eficiencia y el poder de detección de epistasis.
- Los valores *p* solos no permiten ninguna declaración directa sobre la fuerza de la asociación. Un valor *p* solo estima la probabilidad de haber observado el valor del estadístico de prueba bajo la hipótesis nula (es decir, no hay asociación entre el SNP probado y el fenotipo).
- El software BEAM es más exhaustivo que Plink para la búsqueda de interacciones entre SNPs.

4. Glosario

Allec: Alélico

A1: Alelo menor

A2: Alelo mayor

AFF: Recuento de alelos entre casos

BEAM: Bayesian Epistasis Association Mapping

BEST-CHISQ: Estadístico chi cuadrado más grande

BEST-CHR: Cromosoma de mayor variante

BEST-SNP: Identificación mayor variante

BP: Coordenadas de par de base

CC: Caso/control

CHISQ: CHI cuadrado

Chr: Cromosoma

Co: Caso

CMH: Cochran-Mantel Haenzel

DF: Grados de libertad

Dom: Dominancia

FA: Frecuencia de alelos entre casos

FU: Frecuencia de alelos entre controles

GWAS: Estudio de asociación

ID: Identificación del individuo

LD: Desequilibrio de ligamiento

N-sig: Número de resultados de pruebas epistaticas

N-tot: Número de resultados prueba valida

OR_INT: Odds ratio

P: P valor

PRO: Proporción significativa

Rec: Recesivo

STAT: Estadistica chi cuadrado

SNPs: Polimorfismo de un solo nucleótido

Test: Tipos de pruebas

Trend: Cochram –Armitage prueba de tendencia

UNAFF: Recuento de alelos entre controles

5. Bibliografía

- Arkin, Y., Rahmani, E., Kleber, ME, Laaksonen, R., März, W., y Halperin, E. (2014). EPIQ-eficiencia de detección de SNP-SNP epistatic interacciones para rasgos cuantitativos. *Bioinformática*, *30* (12), 19-1925.
- Bateson W. (1909). Mendel's Principles of Heredity. Cambridge, UK: Cambridge University Press.
- Bender D., Maller J., Sklar P., Bakker P. I. W,. Daly M J., and. Sham P C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. vol. 81 (pg. 559-575).
- Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. (1999). Characterization of single nucleotide polymorphisms in coding regions of human genes. Nature Genet. 22:231–238.
- Cordell H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet. 11, 2463–2468.
- Cordell HJ. (2009). Genome-wide association studies: detecting gene-gene interactions that underlie human diseases. Nat Rev Genet.; 10: 392–404.
- Fisher R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edin. 52, 399–433.

- Ferreira T, Donnelly P, Marchini J. (2007).Powerful Bayesian gene-gene interaction analysis. Am J Hum Genet. 81.
- Gayán, J., González-Pérez, A., Bermudo, F., Sáez, M. E., Royo, J. L., Quintas, A., Ruiz, A. (2008). A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. BMC Genomics, 9, 360.
- Herold Christine, Steffens Michael, Felix F. Brockschmidt, Max P. Baur, Tim Becker. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information, Bioinformatics, Volume 25, Issue 24, Pages 3275 3281.
- Hemani G, Theocharidis A, Wei W, Haley C. (2011).EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics. 27(11):1462–1465.
- Jung, J., Sun, B., Kwon, D., Koller, D. L., & Foroud, T. M. (2009). Allelic-Based Gene-Gene Interaction Associated With Quantitative Traits. Genetic Epidemiology, 33(4), 332–343.
- Li, L., Yu, M., Jason, R. D., Shen, C., Azzouz, F., McLeod, H. L., Flockhart, D. A. (2008). A Mixture Model Approach in Gene-Gene and Gene-Environmental Interactions for Binary Phenotypes. Journal of Biopharmaceutical Statistics, 18(6), 1150–1177.
- Mackay TF, Moore JH. (2014). Why epistasis is important for tackling complex human disease genetics. Genome Medicine. 6 (6):42.
- Miller, D. J., Zhang, Y., Yu, G., Liu, Y., Chen, L., Langefeld, C. D., Wang, Y. (2009). An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. Bioinformatics, 25(19), 2478–2485.
- Moore JH, Williams SW. (2002). New strategies for identifying gene-gene interactions in hypertension. Annals of Medicine.34:1–8.
- Moore J. H., Williams S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 27, 637–646.
- Moore J. H., Williams S. M. (2009). Epistasis and its implications for personal genetics. Am. J. Hum. Genet. 85, 309–320. 10.
- Niel C, Sinoquet C, Dina C, Rocheleau G. (2015). A survey about methods dedicated to epistasis detection. Front Genet. 6:285.

- Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. (2002). Functional SNPs in the lymphotoxin-gene that are associated with susceptibility to myocardial infarction. Nature Genetics. 32:650–654.
- Park MY, Hastie T. (2008). Penalized logistic regression for detecting gene interactions. Biostatistics. 9:30–50.
- Prabhu, S., & Pe'er, I. (2012). Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease. Genome Research, 22(11), 2230–2240.
- Purcell S., Neale B., Todd-Brown K., Thomas L, A. R., Ferreira M.,
- Tang W., Wu X., Jiang R., Li Y. (2009). Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLoS Genet.5:e1000464.
- Risch N, Merikangas K. (1996). The future of genetic studies of complex human diseases. Science. 273:1516–1517.
- Ritchie, MD, Hahn, LW, Roodi, N., Bailey, LR, Dupont, WD, Parl, FF y Moore, JH. (2001). Multifactor-Dimensionality Reduction revela las interacciones de alto orden entre los genes de estrógeno-metabolismo en el cáncer de mama esporádico. American Journal of Human Genetics, 69 (1), 138 - 147.
- Rothberg BEG. (2001). Mapping a role for SNPs in drug development. Nature Biotechnology. 19:209–211.
- Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D.(2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409:928–933.
- Schork NJ, Fallin D, Lanchbury JS. (2000). Single nucleotide polymorphisms and the future of genetic epidemiology. Clin. Genet. 58:250–264.
- Siemiatycki J., Thomas D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. Int. J. Epidemiol. 10, 383–387.

- Steen KV. (2012).Travelling the World of Gene-Gene Interactions. Brief Bioinformatics. 13(1):1–19.
- Syvanen AC. (2005). Toward genome-wide SNP genotyping. Nature Genetics. 37:S5–S10.
- Templeton AR. (2000). Epistasis and complex traits. In: Wade M, Brodie B III, Wolf J, editors. Epistasis and Evolutionary Process. Oxford: Oxford University Press.
- Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., & Yu, W. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *American Journal of Human Genetics*, *87*(3), 325–340.
- Wang X., Elston R. C., Zhu X. (2010). The meaning of interaction. Hum. Hered. 70, 269–277.
- Wang, T., Ho, G., Ye, K., Strickler, H., & Elston, R. C. (2009). A Partial Least Square Approach for Modeling Gene-gene and Gene-environment Interactions When Multiple Markers Are Genotyped. Genetic Epidemiology, 33(1), 6–15.
- Wei C, Lu Q. (2014). GWGGI: software for genome-wide gene-gene interaction analysis. BMC Genetics.15:101.
- Wei WH, Hemani G, Haley CS. (2014) Detecting Epistasis in Human Complex Traits. Nat Rev Genet. 15(11):722–33.
- Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics, 25(6), 714–721.
- Zhang YM, Xu S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity.95:96–104.
- Zhang Y, Liu JS. (2007). Bayesian inference of epistatic interactions in case-control studies. Nat Genet.39:1167–1173.
- Zhao J, Xiong M. (2006). Test for interaction between two unlinked loci. Am J Hum Genet. 79:831–845.

6. Anexo

6.1. Códigos y comandos de entrada y salida del programa Plink

6.1.1.Comando de entrada

1. Instalación, Plink versión 1.9 de la página (http://zzz.bwh.harvard.edu/plink/).

2. Plataformas de descarga Linux, MS-DOS, Apple Mac, C/C++.

3. Una vez instalado plink, el archivo ejecutable debe colocarse en el directorio de trabajo actual o en algún lugar de la ruta del comando, utilizando:

Plink o ./plink

4. Para comenzar a ejecutar plink se deben descargar dos archivos con extensión.

.ped y .map

5. Construcción de los archivos con extensión .PED (El archivo PED es un archivo delimitado en espacio en blanco, las primeras seis columnas son obligatorias).

- ID familiar
- Individuo oh identificación de la muestra
- Padre
- Madre
- Sexo

Macho = 1
Hembra = 2
Desconocido = otros

- Fenotipo el estado por defecto, debe codificarse:

-9 desaparecido
0 desaparecidos
1 no afectado
2 afectado

Nota: Si mi archivo está codificado 0/1 para representar no afectado / afectado, entonces se utiliza el siguiente código **--1 flag**:

plink --file mydata –1

Nota: Especificará un fenotipo de enfermedad codificado (El valor de fenotipo faltante para los rasgos cuantitativos es, por defecto, -9 (esto también se puede utilizar para los rasgos de la enfermedad, así como 0). Se puede reiniciar incluyendo la opción.

--missing-phenotype

-9 desaparecido
0 no afectado
1 afectado

● Genotipo (deben ir separados por espacio, tener dos alelos por cada marcador, y si falta se le colocara el numero 0).

Ejemplo entrada del formato de archivos PED:

plink --ped mydata.ped

Example of a PED file of the standard PLINK format:																			
FAM1	NA06985	0	0	1	1	A	T	T	T	G	G	C	C	A	T	T	T	G	G
FAM1	NA06991	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G
0	NA06993	0	0	1	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G
0	NA06994	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G
0	NA07000	0	0	2	1	C	T	T	T	G	G	C	T	C	T	T	T	G	G
0	NA07019	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G
0	NA07022	0	0	2	1	C	T	T	T	G	G	0	0	C	T	T	T	G	G
0	NA07029	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G
FAM2	NA07056	0	0	0	2	C	T	T	T	A	G	C	T	C	T	T	T	A	G
FAM2	NA07345	0	0	1	1	C	T	T	T	G	G	C	C	C	T	T	T	G	G

6. Construcción de los archivos con extensión MAP (Por defecto, cada línea del archivo MAP describe un solo marcador y debe contener exactamente 4 columnas).

● Cromosoma

● Identificación del marcador

● Posición genética

● Posición física

Ejemplo entrada formato archivo MAP:

plink --map mydata.map

Example of a MAP file of the standard PLINK format:			
21	rs11511647	0	26765
X	rs3883674	0	32380
X	rs12218882	0	48172
9	rs10904045	0	48426
9	rs10751931	0	49949
8	rs11252127	0	52087
10	rs12775203	0	52277
8	rs12255619	0	52481

Nota: El archivo MAP debe contener tantos marcadores como el archivo PED. Los marcadores en el archivo PED no necesitan estar en orden genómico: (es decir, el archivo MAP debe alinearse con el orden de los marcadores de archivos PED).

7. Análisis de epistasis: Para las muestras basadas en la población con riesgo de enfermedad, es posible realizar una prueba de epistasis que puede ser solo caso o caso-control. (Estas pruebas de epistasis actualmente solo son aplicables para muestras basadas en la población, no basadas en la familia).

● **Epistasis entre SNPs X SNPs**. (Comandos basados en la población de caso / control).

plink --file mydata –epistasis

Nota: El comando **--epistasis** está configurado para probar una gran cantidad de comparaciones de SNP por SNP, la mayoría de los cuales no serían significativos o de interés. Debido a que la salida puede contener millones o miles de millones de líneas de resultados, el valor predeterminado solo genera pruebas con valores p inferiores a 1e-4, Si el conjunto de datos es mucho más pequeño y definitivamente solo queremos ver todos los resultados, agregue **--epi1 1**. Si no, es probable que se vea un archivo de salida en blanco a excepción del encabezado (es decir, que le indique inmediatamente que ninguna de las pruebas fue significativa en 1e-4).

● **Epistasis solo casos (comando).**

plink --file mydata --fast-epistasis --case-only

Actualmente, en el análisis de solo casos, solo los SNP que están separados por más de 1 Mb, o en diferentes cromosomas, se incluyen en esta prueba. Este comportamiento se puede cambiar con la opción **--gap**, con la distancia especificada kb (Esta opción es importante, ya que la prueba de caso solo para epistasis supone que los dos SNP están en equilibrio de enlace en la población general).

plink --file mydata --fast-epistasis --case-only --gap 5000

● **Interacciones entre Genes. (Análisis estratificado)**

Aquí se pueden realizar una serie de pruebas de asociación de caso / control que toman en cuenta agrupamientos oh pruebas de homogeneidad de efectos de racimo. Utilizando el comando **--within**. (Clúster).

Hay dos clases básicas de prueba:

A) Pruebas para la asociación general enfermedad / gen, controlando por conglomerados.

B) Prueba de heterogeneidad de la enfermedad / asociación génica entre diferentes grupos.

El tipo de estructura de clúster variará en términos de cuántos clústeres hay en la muestra y cuántas personas pertenecen a cada clúster. En un extremo, podríamos tener solo 2 grupos en la muestra, cada uno con una gran cantidad de casos y controles. En el otro extremo, podríamos tener una gran cantidad de clústeres, de modo que cada clúster solo tenga 2 individuos. Estos factores influyen en la elección del análisis estratificado.

Pruebas que se pueden analizar:

● Las pruebas de Cochran-Mantel-Haenszel (CMH) son válidas con una gran cantidad de grupos pequeños y una pequeña cantidad de grupos grandes. Estas pruebas proporcionan una prueba basada en una odds ratio "promedio" que controla la posible confusión debido a la variable del clúster. La prueba CMH viene en dos sabores: 2x2xK e IxJxK. Actualmente, la prueba 2x2xK representa una {enfermedad x SNP clúster} prueba. La forma generalizada, IxJxK, representa una prueba de {clúster x SNP enfermedad}, es decir, el SNP varía entre los grupos, controlando cualquier posible asociación verdadera

de SNP / enfermedad. Esta última prueba podría ser útil para interpretar asociaciones significativas en muestras estratificadas.

● La prueba de Breslow-Day pregunta si los diferentes conglomerados tienen diferentes odds de enfermedad / gen: esta prueba supone un tamaño de muestra moderado dentro de cada grupo. La prueba de asociación total de partición, que es conceptualmente similar a la prueba de Breslow-Day, también hace la misma suposición.

● Comandos para la prueba de Cochran-Mantel-Haenszel.

plink --file mydata --mh --within mycluster.dat

o

plink --file mydata --mh2 --within mycluster.dat

El rango del intervalo de confianza con la opción **--mh** se puede cambiar con la opción **–ci.**

plink --file mydata --mh --within mycluster.dat --ci 0.99

8. Análisis de asociación: La prueba de asociación básica es para un rasgo de enfermedad y se basa en la comparación de las frecuencias de los alelos entre los casos y los controles (se dispone de valores p empíricos y asintóticos).

● Para realizar un análisis de asociación caso / control estándar

plink --file mydata --assoc

● Prueba de Fisher para generar significancia.

plink --file mydata --fisher

● Prueba de asociación de modelos alternativos y completos. (Son pruebas de asociación entre una enfermedad y una variante que no sea la prueba alélica básica)

Las pruebas que se ofrecen son (además de la prueba alélica básica):

● Prueba de tendencia Cochran-Armitage

● Prueba genotípica (2 df)

● Prueba de acción genética dominante (1df)

● Prueba de acción gen recesiva (1df)

Una ventaja de la prueba de Cochran-Armitage es que no supone el equilibrio de Hardy-Weinberg, ya que el individuo, no el alelo, es la unidad de análisis (aunque los valores de p empíricos basados en la permutación de la prueba alélica básica también tienen propiedad). Es importante recordar que los SNP que muestran desviaciones severas de Hardy-Weinberg a menudo son malos SNP, o reflejan la estratificación en la muestra, por lo que probablemente se excluyen mejor en muchos casos.

La prueba genotípica proporciona una prueba general de asociación en la tabla de 2 por 3 de enfermedad por genotipo. Los modelos dominantes y recesivos son pruebas para el alelo menor. El alelo menor que se puede encontrar en el resultado de los comandos **--assoc** o **--freq**. Es decir, D es el alelo menor y d es el alelo principal:

● Comandos a utilizar para estas pruebas:

plink --file mydata --model

● También es posible agregar el indicador **--fisher** para obtener los valores p exactos:

./plink --bfile mydata --model –fisher

Las salidas para los análisis de epistasis las podemos controlar con los siguientes comandos.

plink --file mydata --epistasis --0.0001

6.1.2 Interpretación de las salidas Plink

- ***Salidas de análisis de epistasis***

Extensión plink.epi.cc (caso/control) y plink.epi.co (caso)

Tabla 1: Interpretación de los archivos de salida con extensión plink.epi.cc y plink.epi.co.

plink. epi.cc **plink.epi.co**	**Interpretación de los códigos de salida**

Chr1	Código de cromosoma variante 1
SNP1	Identificador de variante 1
Chr2	Código de cromosoma variante 2
SNP2	Identificador de variante 2
'OR_INT'	Odds ratio (caso / control) o coeficiente de regresión
Stat	Estadístico Chi-cuadrado
p	P-Valor

Extensión plink.epi.cc.summary (caso/control) y plink.epi.co.summary (caso)

Tabla 2. Interpretación de los archivos de salida con extensión plink.epi.cc.summary y plink.epi.co.summary.

Plink.epi.cc.summary **Plink.epi.co.summary**	**Interpretación de los códigos de salida**
Chr	Código cromosómico
SNP	Identificador de variante
N_SIG	Número de resultados de pruebas epistáticas "significativas" (según el valor de ep2)
N_TOT	Número total de resultados de prueba válidos
PRO	Proporción significativa
BEST_CHISQ	Estadística chi-cuadrado más grande
BEST_CHR	Cromosoma de la variante de mayor estadística
BEST_SNP	ID de la variante de mayor estadística

- ***Salidas de análisis de asociación***

Tabla 3. Interpretación del archivo de salida con extensión plink.assoc y plink.fisher.

Plink.assoc / plink.fisher	**Interpretación de los códigos de salida**
Chr	Código cromosómico
SNP	Identificador de variante
BP	Coordenadas de par base

A1	Alelo 1 (generalmente menor)
FA	Alelo 1 frecuencia entre casos
F_U	Frecuencia del alelo 1 entre los controles
A2	Alelo 2
CHISQ	Prueba alélica chi-cuadrado estadística.
p	Prueba alélica p-valor
O	odds (alelo 1 | caso) / odds (alelo 1 | control)

Tabla 4. Interpretación del archivo de salida con extensión plink.model.

Plink.model	**Interpretación de los códigos de salida**
Chr	Código cromosómico
SNP	Identificador de variante
A1	Alelo A1 (generalmente menor)
A2	Alelo A2 (generalmente mayor)
TEST	Tipo de prueba: uno de {'GENO', 'TREND', 'ALLELIC', 'DOM', 'REC'}
AFF	'/' - genotipo separado o recuentos de alelos entre los casos
UNAFF	'/' - genotipo separado o recuentos de alelos entre los controles
CHISQ	Estadística Chi-cuadrado.
DF	Chi-cuadrado grados de libertad.
p	P-valor

6.2. Códigos y comandos de entrada y salida del programa BEAM

6.2.1. Comando de entrada

1. Instalación del programa Beam de la página http://www.mybiosoftware.com/beam-3. disease-association-mapping.html.

2. plataformas de descarga, Linux. El software se escribirá en C ++

3. Una vez descargado BEAM, El archivo ejecutable debe colocarse en el directorio de trabajo actual o en algún lugar de la ruta del comando, utilizando.

./BEAM

4. Formato de datos para la ejecución de BEAM.

● Estado de la enfermedad de cada individuo.

1 Pacientes
0 controles

Ejemplo posición de 3 casos y 3 controles

Id chr position **111000**

● Identificación de SNP, Posición del cromosoma, caso, control

Ejemplo

rs1021 chr5 110123548 1 0 3 0 1 1

SNPs cromosoma posición casos control

Los genotipos deben codificarse como alelos, por ejemplo, 0, 1, 2, que denotan el número de alelos alternativos, y 3 denota datos faltantes. Los alelos faltantes son simplemente imputados por regla general en cada SNPs de forma independiente.

● Modelos y parámetros que mide el programa

✓ **"-filter k":** Deja que el programa filtre los SNPs con demasiados genotipos perdidos alrededor del (3%), falta de equilibrio entre los casos y controles. Si se usa esta opción, se debe especificar el valor k. k = 0 si el heterocigoto está codificado como 2 y k = 1 si el heterocigoto está codificado como 1.

✓ **"-sample burnin mcmc"**: Esta opción especifica los números de burnin y las iteraciones de muestreo. Por defecto, burnin = mcmc = 100. En cada iteración, el programa actualiza todas las variables una vez, y por lo tanto estos números no se relacionan necesariamente con la cantidad de SNPs.

✓ **"-prior p":** Esta opción especifica la probabilidad de que cada SNPs esté asociado con la enfermedad. Por defecto, p = 5 / L, es decir, se esperan 5 SNP asociados (de L SNPs).

✓ **"-T t":** Esta opción ayuda al programa a salirse de los modos locales en las primeras iteraciones.

6.2.2 ***Interpretación de salidas*** BEAM

✓ Esta salida contiene las probabilidades posteriores de asociaciones marginales e interacción por SNPs.

posterior.[outname]

✓ Contiene el gráfico de la enfermedad

g.[outname].dot

✓ Es un archivo de prueba de chisq single-SNP, con estadísticas de prueba y conteos de alelos.

chi.txt