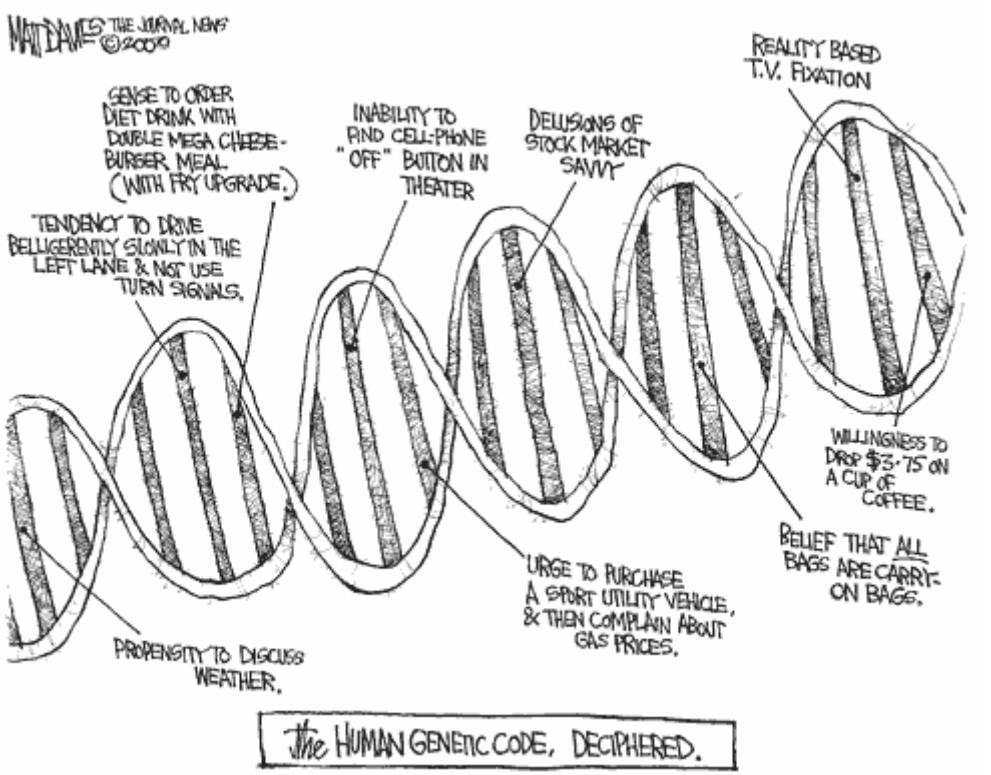
Universidad Autónoma de Madrid
Facultad de Ciencias
Departamento de Biología Molecular

Desarrollo y aplicación de métodos bioinformáticos para el análisis de polimorfismos genéticos

Tesis Doctoral

Lucía Conde Lagoa

Madrid 2007


MATT DAVIES THE JOURNAL NEWS ©2006
REALITY BASED T.V. FIXATION
SENSE TO ORDER DIET DRINK WITH DOUBLE MEGA CHEESE-BURGER MEAL (WITH FRY UPGRADE.)
INABILITY TO FIND CELL-PHONE "OFF" BUTTON IN THEATER
DELUSIONS OF STOCK MARKET SAVVY
TENDENCY TO DRIVE BELLIGERENTLY SLOWLY IN THE LEFT LANE & NOT USE TURN SIGNALS.
WILLINGNESS TO DROP $3.75 ON A CUP OF COFFEE.
BELIEF THAT ALL BAGS ARE CARRY-ON BAGS.
URGE TO PURCHASE A SPORT UTILITY VEHICLE, & THEN COMPLAIN ABOUT GAS PRICES.
PROPENSITY TO DISCUSS WEATHER.
The HUMAN GENETIC CODE, DECIPHERED.

AGRADECIMIENTOS

En primer lugar quisiera expresar mi agradecimiento a Ximo Dopazo, director de esta tesis, por darme la oportunidad de trabajar en su grupo y demostrar así su confianza en mí, y por el esfuerzo y ayuda prestados durante estos años de tesis. Gracias también por el enorme esfuerzo realizado para proporcionarme financiación durante todo este tiempo.

Gracias a todos y cada uno de mis compañeros de grupo, especialmente a los que empezaron conmigo en el CNIO: Javito y Álvaro (gracias por ayudarme en mis primeros pasos en la bioinformática, aun recuerdo ese primer 'ls -l'), Juanma, Pablo (espero tener casa pronto en Japón, te llevaré un tarro de nocilla blanca), Jaime y Leo (¡los mejores compañeros de piso!), Hernán, Ramón y sobre todo a Fátima, la churu, gracias por estar siempre ahí, en los buenos momentos y sobre todo en los malos, ¡¡voy a echar mucho de menos nuestras conversaciones ilustres!! Gracias también a los que llegaron después, en mi etapa en el CIPF, especialmente a David (por estar siempre dispuesto a escuchar mis historias y por tus consejos), Jordi (por tu enoooorme paciencia, ¡aun te debo una cena!), Eva (te quedas sola en el gineceo), Toni (quien tuvo el honor y la mala suerte de ayudarme con las correcciones), Joaquín y Nacho (ahora te toca seguir a ti), y a los que me conocieron sentada en el ordenador escribiendo esta tesis, Marc, Stefan, Ana, Emidio y Peio. No me olvido de David Casado, Poyatos, Santoyo y de los que pasaron en algún momento por el grupo y me dejaron buenos recuerdos. Gracias a todos, por vuestra ayuda, por los buenos ratos dentro y fuera del labo, porque esto nunca hubiera sido lo mismo sin vosotros.

Gracias a Impa y a Fon (a quienes veo menos de lo que quisiera), por hacerme siempre un hueco cuando vuelvo a Pontevedra; a la Shamber, con la que me he reído tela; a Ero y Chisco, mi comunidad gallega en Madrid; a Martita y a Vane Castro, mis amigas desde la Uni; a Inma y Ros, las paquitas. Gracias a Blanca. Gracias a Imelda, el Seco y Rafa, por su cachondeo y por enseñarme Valencia de noche ;-). Gracias a Eran Halperin, por todo lo que me ha enseñado y por el tiempo que me dedicó cuando estuve en ICSI. Gracias también a Érika y René, porque a pesar de estar a 10,000 kilómetros me hicieron sentir como si estuviera en casa. Gracias a Jimmy, por su increíble ayuda con la estadística y el inglés, por llevarme a los mejores tailandeses, por los viajes que hemos hecho y espero seguir haciendo, por su paciencia y en definitiva por ser como es. Gracias, porque entre todos habéis hecho que mi paso por Madrid, Valencia y Berkeley haya sido muy gratificante.

Gracias a los miembros del tribunal, por permitirme contar con los mejores científicos en mi lectura de tesis.

Gracias a mis hermanos David y Santi. A mi sobri Dani.

Finalmente gracias a mis padres, a quienes dedico esta tesis, por la educación y apoyo que siempre me han dado. Mi agradecimiento hacia vosotros es impagable.

ÍNDICE

ABREVIATURAS

aCGH	*Array* de hibridación genómica comparativa, *comparative genomic hybridization array*
API	Interfaz de programación de aplicaciones, *application programming interface*
BAC	Cromosoma artificial bacteriano, *bacterial artificial chromosome*
BRE	elemento de reconocimiento del factor TFIIB, *TFIIB recognition element*
CD/CV	Enfermedad común/variación común, *common disease/common variation*
CNV	Variación en el número de copia, *copy number variation*
CGH	Hibridación genómica comparativa, *comparative genomic hybridization*
CPM	*Combinatorial Partitioning Method*
DAS	Sistema de anotación distribuida, *distributed annotation system*
DPE	Elemento promotor río abajo, *downstream promoter element*
EM	*Expectation Maximization*
ESE	Potenciador de *splicing* exónico, *exonic splicing enhancer*
ESS	Silenciador de *splicing* exónico, *exonic splicing silencer*
GPNN	*Genetic Programming Optimized Neural Network*
HGMD	*The Human Gene Mutation Database*
HWE	Equilibrio de Hardy-Weinberg, *Hardy-Weinberg equilibrium*
Inr	Elemento iniciador, *initiator element*
LD	Desequilibrio de ligamiento, *linkage disequilibrium*
LOD	Logaritmo de disparidad, *logarithm of odds*
MAF	Frecuencia del alelo minoritario, *minor allele frequency*
MDR	*Multifactor Dimensionality Reduction*
NCBI	*National Center for Biotechnology Information*
NMD	Degradación mediada por mutaciones terminadoras, *nonsense-mediated mRNA decay*
nsSNP	SNP codificante no sinónimo, *non-synonymous SNP*
OMIM	*Online Mendelian Inheritance in Man*
PAC	Cromosoma artificial derivado del fago P1, *P1-derived artificial chromosome*
PAM40	*Point Accepted Mutation-40*
PAML	*Phylogenetic Analysis by Maximum Likelihood*
PCR	Reacción en cadena de la polimerasa, *polymerase chain reaction*
PSAT	*Population Stratification Association Test*
PSSM	*Position Specific Scoring Matrix*
PWM	Matriz de pesos de posiciones, *position weight matrix*
RF	*Random Forests*
RFLP	Polimorfismo de longitud de fragmentos de restricción, *restriction fragment length polymorphism*
RNPnh	Ribonucleoproteína nuclear heterogénea

RPM	*Restricted Partition Method*
SAA	*Set Association Approach*
SELEX	S*ystematic Evolution of Ligands by Exponential Enrichment*
SLR	*Sitewise Likelihood-Ratio*
SNP	Polimorfismo de un solo nucleótido, *single nucleotide polymorphism*
SR	Proteínas ricas en serina/arginina
TF	Factor de transcripción, *transcription factor*
TFBS	Sitio de unión a factor de transcripción, *transcription factor binding site*
TTS	Secuencia capaz de formar triple hélices, *triplex-forming oligonucleotide target sequence*
TSS	Sitio de inicio de la transcripción, *transcription start site*
WTCCC	*The Wellcome Trust Case Control Consortium*

SUMMARY

With the completion of the sequencing of the human genome, much attention has been centered on the study of human genome variability. Single nucleotide polymorphisms (SNPs) are the most common source of human genetic variation, and they are, undoubtedly, a valuable resource for investigating the genetic basis of diseases. SNPs, together with DNA copy number variations (CNVs), have become one of the most actively researched areas of genomics in recent years.

Although the majority of these variations probably results in neutral phenotypic outcomes, certain polymorphisms can predispose individuals to disease, or influence its severity or progression. One of the biggest challenges in biomedical research is the identification of these variants, which are usually prioritized for their inclusion in association studies. The recent ability to collect a large number of SNPs for a given individual, has led researchers to conduct large scale association studies with varying disease outcomes. These studies have become a powerful tool for the investigation into the association between genetic variation and disease. Much care is needed when conducting such studies; they require a careful process of study design, analysis and interpretation of data, and an intelligent application of bioinformatics methods is essential.

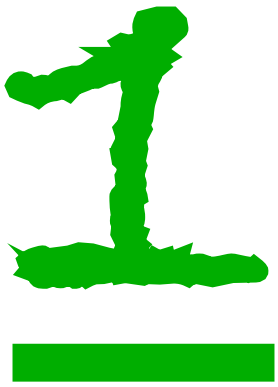
In this thesis, novel bioinformatics methods are introduced to facilitate the analysis of genetic polymorphisms. The methods have been designed and documented around relevant tools to aid the scientific community analyze these data without the typical hurdles met when dealing with the complexity generally encountered in this field. In the same way that tools like blast have facilitated interesting research in novel areas, the tools we describe here aim to provide leverage to researchers in the same manner; by freeing researchers from the difficult task of tool development, more productive downstream research can occur.

First, methods are developed which aid in the prediction of the functional impact of SNPs. While much attention has been focused on the effects of variation on the amino acid sequence, variations that disrupt gene regulation, expression or splicing can dramatically impact gene function. This work approaches for first time the possible effect of these regulatory variations. All the methods have been implemented in a software suite, PupaSuite (http://pupasuite.bioinfo.cipf.es), which is part of the Centro Nacional de Genotipado for SNP selection in association studies.

Second, a tool for visualization and analysis of array CGH data has been developed with the purpose of studying genomic copy number variations (ISACGH, http://isacgh.bioinfo.cipf.es). In addition to identification of the genomic regions which contain altered copy number, the tool allows one to analyze the relationships of CNVs to gene expression changes and to functional annotation within relevant regions. The combined information produced using ISACGH can aid the study and

interpretation of phenotype in the context of array CGH data.

Finally, a method is developed for analysis of genotype data in combination with functional information, with the aim of finding functionally related polymorphisms, which, as a block, present high association to the phenotype or disease. The method proposed helps to avoid many of the common statistical problems encountered in this setting.



INTRODUCCIÓN

1. Variación genética

Las secuencias completas de dos genomas humanos son por término medio un 99.9% idénticas (www.wellcome.ac.uk/genome). El 0.1% restante varía entre cada individuo, siendo las variaciones más comunes las que se conocen como SNPs o polimorfismos de una sola base.

En el primer análisis del genoma humano se detectaron aproximadamente 1.42 millones de SNPs (Lander *et al.*, 2001; Venter *et al.*, 2001). Hoy en día aparecen descritos 11 millones de SNPs en la base de datos dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP) del National Center for Biotechnology Information de Estados Unidos (NCBI), un desarrollo que se ha acelerado a partir de 1999 gracias a la formación del consorcio SNP y más adelante se ha consolidado con el proyecto HapMap (International HapMap Consortium, 2005) (figura 1).

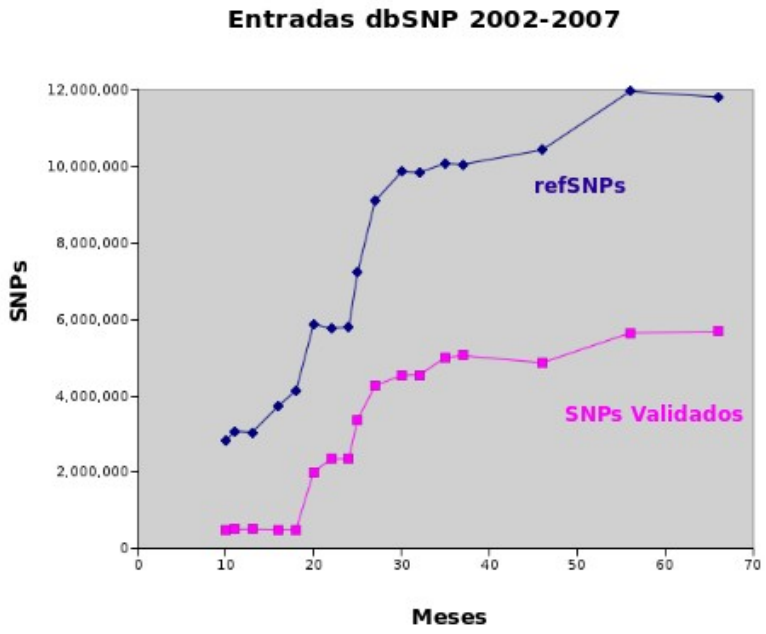


***Figura 1**. Crecimiento en el número de SNPs incluidos en la base de datos dbSNP. En agosto de 2002 había descritos 2,817,196 SNPs, y menos de una cuarta parte de ellos estaban validados. En marzo de 2007 la cifra ha aumentado a 11,811,594 SNPs en total, con 5,689,286 (48.16%) de ellos validados. Fuente dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP).*

Un SNP es una variación de una sola base en una determinada posición del DNA genómico, en la que el alelo (una de las formas variantes del SNP) menos frecuente tiene al menos una abundancia de un 1% en la población (Brookes, 1999). Un SNP puede ser una deleción, inserción o sustitución de una base, y teóricamente puede tener hasta 4 alelos, pero más de dos ocurre con frecuencia muy baja.

Normalmente los SNPs son sustituciones de una base, y si la mutación se fija en la población, existirán dos alelos, el normal y el mutado. Para que aparezcan un tercer y un cuarto alelo el mismo nucleótido tiene que mutar otra vez en un individuo y el tercer alelo debe fijarse también en la población. La combinación entre los beneficios adaptativos causados por esas mutaciones, y la selección natural, así como la deriva genética, moldean el genoma en patrones únicos de variaciones genéticas en distintas regiones.

Los SNPs aparecen en cualquier parte del genoma, pero el análisis de la distribución de los polimorfismos a lo largo del genoma humano muestra variaciones significativas en la densidad de polimorfismos y en la distribución de las frecuencias alélicas. Chakravarti (Chakravarti, 1999) mostró una diferencia entre la densidad de SNPs en regiones exónicas y en regiones intergénicas o intrónicas, ya que aparecen en intervalos medios de 1.2Kb en las primeras y en intervalos de 0.9Kb en las segundas.

Atendiendo a sus posibles efectos funcionales, las variaciones se pueden subdividir en distintas clases. Si un SNP está localizado en una región codificante, la variación puede resultar en un cambio de aminoácido y alterar la secuencia de la proteína (SNP no-sinónimo). El SNP también puede ser funcional aunque no produzca ningún cambio de aminoácido (SNP sinónimo), ya que pueden alterar la estabilidad del mRNA (Capon *et al.*, 2006). Los SNPs en las regiones intergénicas o intrones también pueden ser funcionales si alteran los sitios de empalme (*splicing),* o si interrumpen o crean nuevos sitios de unión a factores de transcripción o sitios que actúan como potenciadores o silenciadores de la transcripción.

En enfermedades complejas, probablemente lo más común sean los cambios sutiles como las sustituciones sinónimas y SNPs en zonas intergénicas e intrones, donde la mutación sólo aumenta la susceptibilidad a la enfermedad pero no causa directamente la enfermedad.

SNPs como marcadores genéticos.

Los avances en tecnología molecular conseguidos en los últimos 25 años, han proporcionado un número mayor de marcadores genéticos de forma cada vez más económica (Elston y Spence, 2006).

En 1980 (Botstein *et al.*, 1980) se propusieron los polimorfismos de longitud de fragmentos de restricción (RFLPs) como marcadores para el escaneo completo del genoma. Mediante análisis de ligamiento se estudiaban los patrones de cosegregacion de los marcadores en familias para la localización de genes. Posteriormente, después del desarrollo de la tecnología de la reacción en cadena de la polimerasa (PCR) (Mullis *et al.*, 1986, 1992), los marcadores elegidos fueron los

microsatélites (Weber y May, 1989), en los cuales las sondas utilizadas son secuencias cortas de DNA más fáciles de obtener que las que se necesitan con RFLPs.

Hoy en día, los SNPs se han convertido en los marcadores de elección. Con la tecnología actual es posible genotipar cientos de miles de SNPs por unos pocos dólares por individuo. Además, con la finalización del proyecto HapMap (International HapMap Consortium, 2005) ahora se tienen identificados la mayoría de SNPs en sus localizaciones en el genoma humano. Ésto, unido al hecho de que los SNPs son muy abundantes en el genoma y relativamente estables a lo largo del tiempo, los hace muy útiles como marcadores en estudios genéticos, especialmente en análisis de ligamiento y estudios de asociación (Kruglyak, 1999; Risch y Merikangas, 1996).

Variaciones en número de copia (CNVs)

Los SNPs son probablemente la variación genómica mas común, sin embargo no es el único tipo de variación genómica en humanos. La finalización de la secuencia consenso del genoma humano y el desarrollo de nuevas tecnologías para detectar la posición y la extensión de las alteraciones genómicas han mostrado que existen fragmentos grandes del genoma que han sido delecionados o duplicados (Iafrate *et al*., 2004; Sebat *et al*., 2004). Estos reordenamientos genómicos pueden cambiar el número de copia de los genes situados en esas regiones y por tanto alterar la regulación génica (Lee y Lupski, 2006).

En 2006 (Redon *et al*., 2006) se publicó el primer mapa de CNVs en el genoma humano, en el que se identificaron CNVs en 1400 regiones que solapan con un 14.5% de los genes implicados en enfermedades humanas listados en OMIM (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM). Al igual que con los SNPs, es probable que la mayoría de CNVs sean variantes benignas que no causen enfermedad, sin embargo, hay variaciones específicas asociadas con enfermedades comunes mendelianas (Sebat *et al.*, 2007) y se han descrito diferentes condiciones, incluyendo Parkinson, Alzheimer o la enfermedad de Crohn, cuya susceptibilidad a desarrollarse podría estar influenciada por CNVs (Lee y Lupski, 2006).

Las aproximaciones que clásicamente se han utilizado para el estudio de estas aberraciones genéticas utilizan la hibridación genómica comparativa (CGH), donde el DNA genómico se hibrida con cromosomas en metafase (Kallioniemi *et al*., 1992). Sin embargo, con el reciente desarrollo de las tecnologías de los *microarrays* de DNA ahora es posible el estudio de CNVs a través de estas aproximaciones genómicas masivas mediante el uso de los llamados *arrays* de CGH (aCGH). Con esta técnica se sustituyen los cromosomas en metafase por clones mapeados de forma precisa en el

genoma y colocados en el *microarray* de forma automatizada. Los DNAs de muestra y de referencia se marcan con diferentes sondas flourescentes y se hibridan conjuntamente en el *microarray.* El ratio de flourescencia resultante se mide, clon a clon, y se representa en sus respectivas localizaciones genómicas (figura 2). La resolución para detectar CNVs viene dada por el número y tamaño de los clones del array, los cuales pueden ser oligonucleotidos (25-80bp), cDNA (0.5-2Kb) o insertos de DNA genómico (hasta 200Kb). En los últimos anos se han desarrollado diferentes herramientas bioinformáticas para el análisis de los datos procedentes de aCGHs, así como para la interpretación biológica de los resultados. Entre ellas están CAPweb (Liva *et al*., 2006), ArrayCyGHt (Kim *et al*., 2005) y ISACGH (Conde *et al.*, 2007a, 2007b).

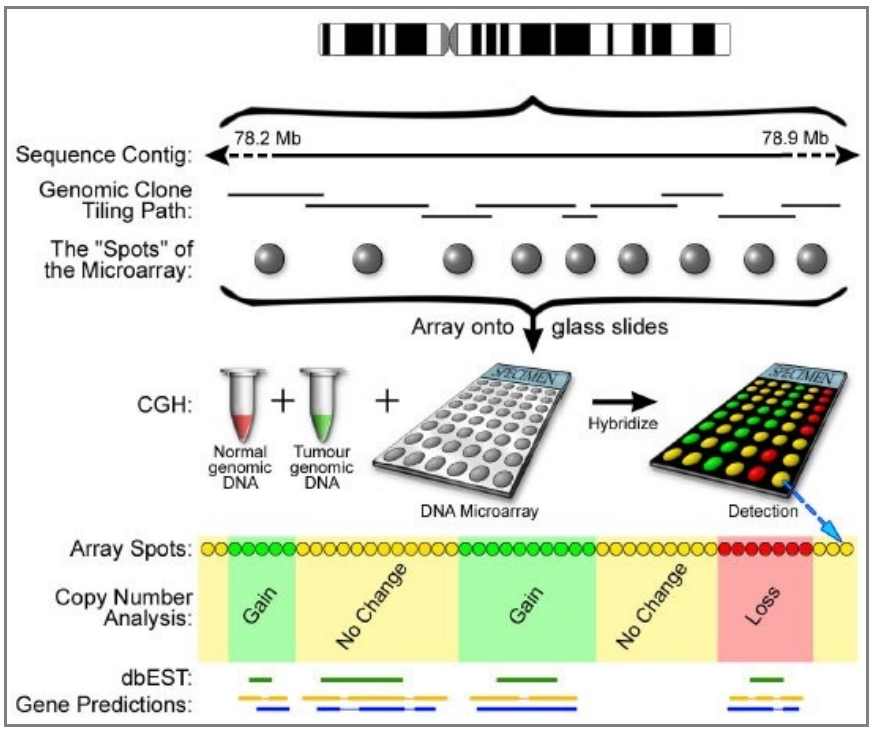


Figura 2*. Representación esquemática de la técnica de arrays de CGH. Se generan arrays de clones genómicos (BACS, PACS, cósmidos...) para cubrir la región de interés. Después de la extracción y purificación, esas secuencias se colocan en las placas del array. Se hibrida la muestra de DNA genómico normal (marcada con Cy3) y la muestra problema, por ejemplo cancerosa (marcada con Cy5), y se detectan las señales con un escáner. Cada punto del array se alinea de forma contigua por sus posiciones cromosómicas y se analiza el ratio de fluorescencia para identificar las regiones de cambio de número de copia. Los resultados se pueden correlacionar con técnicas in silico para identificar genes de interés candidatos. (Figura obtenida de Beheshti et al., 2002).*

Aunque el estudio de las CNVs es relativamente reciente, se están empezando a tener muy en cuenta, y los estudios genéticos empezarán también a incorporar una evaluación de las CNVs en la población de estudio para determinar si es una CNV individual, y no un SNP, el responsable del rasgo que se está estudiando (Lupski, 2007).

2. Estudios genéticos

2.1. Análisis de ligamiento y estudios de asociación

Gran parte de nuestro conocimiento actual sobre la relación entre genotipo y enfermedad proviene de estudios estadísticos donde se mide la correlación entre determinadas variantes genéticas y la probabilidad de desarrollar una enfermedad específica.

Análisis de ligamiento

Los análisis de ligamiento, donde se sigue el rastro del patrón de trasmisión de marcadores genéticos dentro de una familia, han tenido mucho éxito en la identificación de más de un millar de genes de enfermedades monogénicas humanas (Botstein y Risch, 2003).

En el caso de enfermedades complejas comunes como hipertensión, asma o cáncer, no ha habido sin embargo tanto éxito (Altmuller *et al.*, 2001), aunque unas notables excepciones son la implicación de APOE en la enfermedad de Alzheimer (Strittmatter y Roses, 1996) y el papel de NOD2 en la enfermedad de Crohn (Hugot *et al.*, 2001).

La dificultad en trazar el origen genético de estas enfermedades proviene de que la susceptibilidad a desarrollar las mismas depende generalmente de un efecto combinado de muchos polimorfismos en varios genes, a menudo combinado con factores medioambientales. El riesgo de una variante genética sola es pequeño, tanto que están por debajo del límite de detección por análisis de ligamiento donde las muestras son normalmente demasiado pequeñas para proporcionar una significación estadística en la relación enfermedad/genotipo.

Estudios de asociación

Los estudios de asociación, basados en el análisis de diferencias genéticas, particularmente SNPs, entre casos y controles en una población más amplia, son más poderosos para detectar este tipo de señales pequeñas.

En los estudios de asociación normalmente se analizan polimorfismos de los genes de interés o

polimorfismos cercanos. También pueden hacerse estudios de asociación en regiones candidatas o en genomas enteros, y normalmente se utilizan cuando los análisis de ligamiento no pueden dar más información. El objetivo es investigar variaciones en el gen candidato para determinar si un alelo específico o un genotipo está asociado a un mayor riesgo de tener la enfermedad. Esto puede hacerse o bien estudiando las variaciones funcionales del gen o bien estudiando las variaciones genéticas que no son directamente las causantes pero que están asociadas al alelo responsable desconocido. Esta última aproximación es la más común y está basada en un fenómeno genético llamado desequilibrio de ligamiento.

2.2.Desequilibrio de ligamiento (LD)

No todos los alelos se heredan de forma independiente, ya que su proximidad en el cromosoma hace que en una determinada frecuencia se hereden de forma conjunta. Así, alelos de distintos loci a veces tienden a heredarse juntos más a menudo de lo esperado por azar. Estos alelos se dice que están en desequilibrio de ligamiento (LD). Imaginemos que tenemos dos variaciones próximas, una C/T y otra G/T, donde la frecuencia del alelo C es f(C) y la de G es f(G). Si existe segregación al azar entre los dos loci, se esperaría encontrar todas las combinaciones posibles de los cuatro alelos en la población (C-G, C-T, T-G y T-T) con frecuencias que dependerán de las frecuencias individuales de los alelos. El alelo C podría heredarse junto con el alelo G con una frecuencia:

$$f(CG)=f(C)*f(G)$$

Si esto no es así entonces se dice que los alelos C y G están en LD. El grado de LD puede variar entre dos extremos: no LD (segregación al azar) o LD completo. Si existe LD completo entre los alelos anteriores entonces esperaríamos ver a C y G segregando siempre juntos. En este caso sólo habría dos combinaciones de alelos presentes en la población (C-G y T-T).

Existen varias medidas de LD. Las dos más comunes en la literatura son el coeficiente de Lewontin D' y el coeficiente de correlación r^2 (Devlin y Risch, 1995). Éste último refleja el poder estadístico para detectar el LD; el r^2 entre un marcador y un SNP causativo proporciona el tamaño de muestra que se requeriría para detectar la asociación con la enfermedad cuando se genotipa directamente el SNP causativo, en comparación con el que se necesitaría para obtener el mismo poder estadístico

genotipando el marcador. En otras palabras, r^2 mide el poder estadístico que un SNP '1' tiene para predecir los genotipos de otro SNP '2' y viceversa.

Cuando los valores de D' y r^2 difieren significativamente de 0, entonces es que hay evidencia de LD, y un valor de 1 indica LD completo.

Cuando aparece una nueva mutación y se extiende a la siguiente generación, el alelo mutado se heredará junto con los alelos de alrededor. Estos alelos estarán entonces en LD con los otros, y el bloque de alelos coheredados constituirá un haplotipo.

El haplotipo y el LD entre alelos puede romperse por eventos de recombinación. En las primeras generaciones después de la nueva mutación, la recombinación separará con más probabilidad alelos que estén más separados en el genoma, pero después de muchas generaciones separará incluso alelos muy cercanos. Por tanto, es esperable que el LD entre alelos raros, que son relativamente jóvenes, sea mayor que el LD entre alelos comunes (Reich *et al.*, 2001). Por la misma razón, el LD disminuye con la distancia entre marcadores. El grado de LD entre dos alelos puede verse afectado también por otros procesos como nuevas mutaciones, selección, conversión génica, deriva genética y mezclas de poblaciones (Ardlie *et al.*, 2002). Ésto hace que el LD a lo largo del genoma humano sea altamente variable (Dawson *et al.*, 2002). Así, hay regiones cromosómicas de hasta 550 Kb en las que se detecta LD y otras regiones con ningún o poco LD entre marcadores separados por sólo unas pocas Kb (Goddard *et al.*, 2000). Además el grado de LD no solo varía entre regiones cromosómicas, sino también entre distintas poblaciones (Reich *et al.*, 2001).

LD en análisis de ligamiento y estudios de asociación

En su nivel mÁs básico, la asociación genética y los análisis de ligamiento se basan en principios y suposiciones similares (Borecki y Suárez, 2001). Ambos se basan en el hecho de que alelos en loci próximos al locus asociado a una enfermedad tienden a segregar juntos (Hoh y Ott, 2003). En ausencia de entrecruzamientos, el cromosoma que lleva el locus de enfermedad y los alelos de otros loci del mismo cromosoma se transmitirán como bloque (haplotipo).

Los análisis de ligamiento se centran en identificar haplotipos que se heredan intactos en familias o pedigríes de ancestro conocido, y en cambio la asociación se basa en la retención de variantes genéticas adyacentes a través de varias generaciones (Cardon y Bell, 2001). Así, los estudios de asociación pueden ser considerados como estudios de ligamiento muy grandes de pedigríes hipotéticos no observados.

En poblaciones crecientes, como humanos, la recombinación es la principal fuerza que elimina el

ligamiento y la asociación a lo largo de las generaciones (Slatkin, 1994). Cuando aparece una mutación funcional – tal vez una que contribuye a la enfermedad – lo hace en un haplotipo de otras variantes génicas que ya existían. Debido a que el ligamiento se centra sólo en ancestros recientes y normalmente observables, en donde ha habido relativamente pocas oportunidades para la recombinación, las regiones del gen de la enfermedad que se identifican por ligamiento serán a menudo grandes, y pueden incluir cientos o incluso miles de genes a lo largo de muchas megabases de DNA. Por el contrario los estudios de asociación surgen de recombinación histórica, así que las regiones asociadas a la enfermedad son en teoría mucho más pequeñas en poblaciones de apareamientos al azar (Hartl y Clark, 1997), incluyendo sólo un gen o un fragmento génico, siendo por tanto una técnica que permite refinar la localización de un gen.

A lo largo de generaciones subsecuentes, a medida que la mutación se transmite, la recombinación hará que se separe de los alelos específicos de su haplotipo original. Otras variaciones genéticas podrán permanecer juntas en haplotipos ancestrales a lo largo de muchas generaciones, en LD.

Hay varias ventajas de los estudios de asociación frente a los análisis de ligamiento. La primera es que tienen una mayor precisión para localizar el locus de susceptibilidad y particularmente para genes con pequeños efectos individuales, por lo que son de más utilidad que el análisis de ligamiento en el caso de enfermedades complejas (Risch y Merikangas, 2006).

Además, los análisis de ligamiento, que tradicionalmente es el método genético más fiable para enfermedades mendelianas, ha resultado ser mucho menos fiable en el estudio de enfermedades no mendelianas, dando una tasa muy alta de falsos positivos (Risch, 2000). Finalmente, los estudios de asociación se pueden hacer con grupos de individuos no relacionados, simplificando así el proceso de reclutamiento y por tanto haciendo posible que se puedan estudiar muestras poblacionales más grandes.

2.3.El proyecto HapMap

A pesar de la aparente complejidad de los patrones de LD en el genoma, diversos estudios han propuesto que dicho patrón se puede describir como una estructura formada por series de bloques de haplotipos, es decir, de grupos de variantes génicas próximas en el genoma que se heredan juntas y donde la recombinación es rara o ausente.

Este modelo de estructura en bloques de haplotipos tiene importantes aplicaciones en los estudios

de asociación. En respuesta a estos estudios, en octubre de 2002 el United States National Human Genome Research Institute (NHGRI) lanzó una iniciativa llamada proyecto internacional HapMap con el objetivo de caracterizar patrones de LD y haplotipos a lo largo del genoma humano, así como para identificar subgrupos de SNPs (tag SNPs) que capturen la mayoría de la información sobre estos patrones y permitir estudios de asociación genética a gran escala (Collins *et al.*, 2003). Con esta iniciativa se podría minimizar el problema de genotipado de genomas completos al poder reducir el número de SNPs necesarios para genotipar un individuo de 10 millones a tan sólo 500,000 tag SNPs (figura 3).

El proyecto HapMap es un esfuerzo común entre numerosos centros de investigación de distintos países en los que se emplea un amplio rango de plataformas tecnológicas, y tanto los datos y estadísticas como las herramientas necesarias para acceder a los datos son públicos (http://www.hapmap.org).

El primer objetivo del proyecto fue genotipar un SNP cada 5Kb a lo largo del genoma humano (600,000 en total), priorizando la inclusión de SNPs codificantes y SNPs con una frecuencia del alelo minoritario (MAF) mayor del 5%. El genotipado se realizó sobre 270 individuos de 4 poblaciones diferentes. También, y como parte del proyecto ENCODE (ENCODE Project Consortium, 2004) donde se pretende caracterizar todos los elementos funcionales presentes en un pequeño porcentaje del genoma humano, se secuenciaron 10 regiones de 500Kb en 48 individuos no relacionados y todos los SNPs en esas regiones se genotiparon en las 4 poblaciones anteriores.

En marzo de 2005 se ha completado la llamada fase I del proyecto y los resultados mejoraron las expectativas iniciales, ya que no sólo genotiparon 600,000 SNPs sino que se llegó a 1 millón de SNPs genotipados. En la fase II del proyecto el objetivo es aumentar la densidad de SNPs genotipados con 4.6 millones de nuevos SNPs (International HapMap Consortium, 2005).

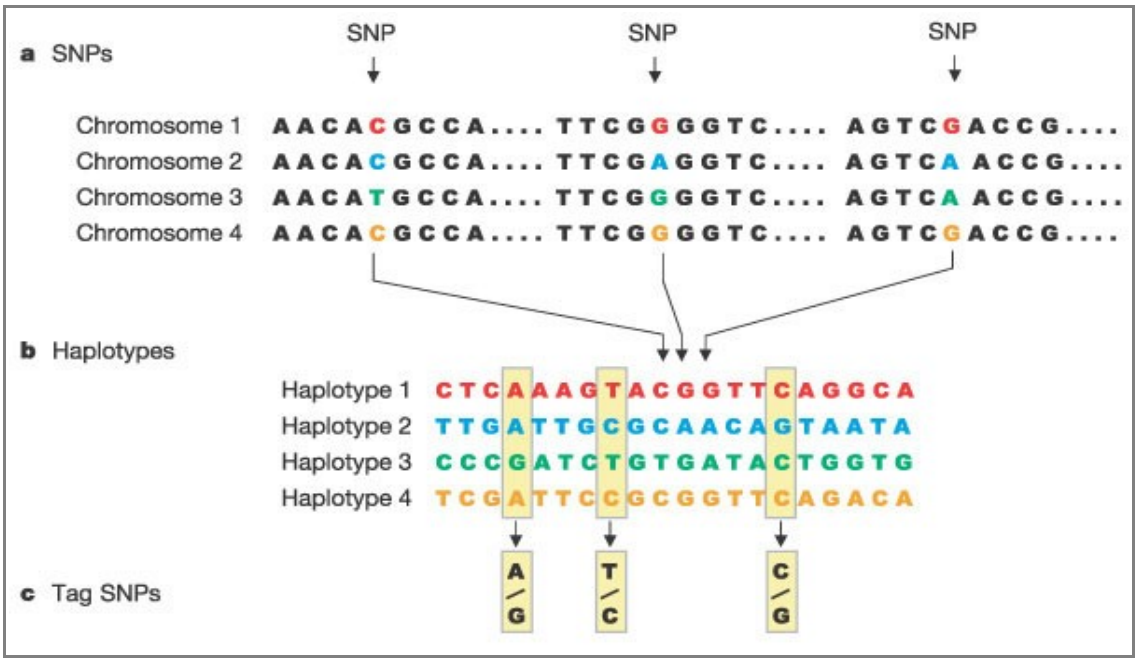


Figura 3. a) *Los SNPs se identifican en muestras de DNA de varios individuos.* ***b)*** *SNPs contiguos que se heredan conjuntamente forman los haplotipos.* ***c)*** *Los "Tag" SNPs en los haplotipos identifican inequívocamente esos haplotipos. Si se genotipan los tres tag SNPs, se podrían identificar cual de los 4 haplotipos de la figura está presente en cada individuo (http://www.hapmap.org/whatishapmap.html).*

El proyecto proporciona información muy útil para el mapeo genético de enfermedades donde la mutación causal real se crea que es una mutación común, ya que HapMap se ha centrado en mutaciones comunes. De acuerdo con varios autores (Weiss y Clark, 2002; Hirschhorn y Daly, 2005), la aproximación de utilizar bloques de haplotipos y tag SNPs para estudios de asociación no será particularmente poderosa para enfermedades causadas por variaciones raras, ya que los marcadores seleccionados por LD en HapMap no podrán marcar los alelos de enfermedad de forma precisa. Por el contrario, el consorcio internacional HapMap dice que incluso las variaciones raras pueden ser descubiertas utilizando este método ya que las variantes poco comunes pueden ser parte de los haplotipos comunes (International HapMap Consortium, 2005). Esto se debe a que la mayoría de los alelos raros probablemente hayan surgido recientemente, y por tanto, es improbable que eventos de recombinación o mutaciones hayan podido interrumpir el haplotipo en el que han aparecido.

Esta polémica que rodea a HapMap surge de la existencia de dos hipótesis alternativas sobre la naturaleza de los alelos que influyen en enfermedades comunes (Chakravarti, 1999). Una de dichas hipótesis sostiene que los alelos son raros y específicos de población (Pritchard, 2001), mientras que la otra hipótesis, conocida como la hipótesis de enfermedad común/variante común (CD/CV) (Reich y

Lander, 2001) propone que hay un número limitado de alelos relativamente frecuentes y que cada uno confiere un riesgo moderado de susceptibilidad a la enfermedad. Algunas observaciones parecen apoyar esta hipótesis, como la de que algunos alelos que ahora predisponen a la enfermedad pudieron ser ventajosos en el pasado, como los alelos que favorecen la acumulación de grasa, y, por tanto, predisponen a la obesidad. Además es probable que la presión selectiva sea débil en enfermedades de aparición tardía y en variantes que contribuyen sólo con un riesgo moderado (Hirschhorn y Daly, 2005; Balding, 2006). La hipótesis de CD/CV es quizá la más ampliamente aceptada, y aunque no es universalmente cierta, lo cierto es que con los datos de HapMap se pueden encontrar alelos comunes que influyen en enfermedades comunes, a pesar de que sean sólo una fracción del total de alelos que influyen en la enfermedad. La identificación de variantes raras requerirá diferentes aproximaciones, como la secuenciación directa de genes candidatos.

2.4. Análisis de casos y controles

Un estudio de asociación se puede llevar a cabo mediante el llamado análisis de casos y controles, en el cual los genotipos de pacientes no relacionados se comparan con los de controles sanos. La medida estadística de asociación en una cohorte de casos/controles puede ser un simple test de Chi cuadrado (χ^2), donde se comparan las frecuencias alélicas entre dos grupos.

Uno de los principales problemas de este tipo de análisis es encontrar los controles adecuados. Por ejemplo, si la asociación genética es una diferencia genética dependiente del sexo, solo podrá verse si se utilizan exclusivamente hembras en el estudio. Para enfermedades de aparición tardía, la edad de los controles es muy importante. Hay un riesgo de perder asociación si la población control es más joven que el grupo de casos, ya que varios de los controles pueden llevar el alelo de la enfermedad y desarrollarla más tarde.

La presencia de artefactos se puede explorar si se comprueba que las frecuencias alélicas entre controles satisfacen el equilibrio de Hardy-Weinberg (HWE). El HWE es el estadístico que determina qué frecuencias deben de observarse en la población para cada genotipo en función de las frecuencias observadas de los alelos para cada locus. Si no se cumple, es indicativo de un posible problema en la selección o análisis de los individuos control y podría motivar la invalidación del estudio de asociación (Campbell y Rudan, 2002).

Para todos los tipos de estudios de asociación es importante tener un grupo de controles de la

misma población que los casos, o existirá un riesgo elevado de tener falsos positivos, que surgen cuando cuando los casos están más relacionados entre sí que con los controles: si entre los casos hay sobrepresentada una población o subgrupo genético, cualquier SNP con diferentes frecuencias alélicas en el subgrupo y en la población general será asociado erróneamente con una de las condiciones (o casos o controles). Estas falsas asociaciones ocurren cuando existe estratificación de población en las muestras.

Estratificación de poblaciones

Una de las principales dificultades cuando se trazan inferencias causales a partir de estudios de asociación de casos y controles es el efecto de la estratificación. Se dice que existe estratificación en la población estudiada cuando ésta no es genéticamente homogénea sino que contiene una mezcla de individuos que pueden ser separados en otros rasgos étnicos distintos al fenotipo investigado. En estos casos las diferencias en frecuencias alélicas entre casos y controles pueden deberse a diferencias sistemáticas en la ascendencia en vez de a la asociación de genes con la enfermedad (Freedman *et al*., 2004; Clayton *et al*., 2005), dando lugar a conclusiones erróneas. Estas falsas asociaciones se conocen como errores de tipo I, donde se rechaza la hipótesis nula de no asociación cuando es de hecho correcta y no hay asociación verdadera. Por tanto la elección de individuos para el estudio de asociación tiene que hacerse con mucha cautela para asegurarse una población homogénea y así evitar la estratificación.

Existen varios métodos que permiten controlar la estratificación. Uno de ellos, el test de transmisión de desequilibrio (TDT) es un diseño de casos/controles que utiliza controles familiares cuando los genotipos parentales se conocen (Spielman *et al.*, 1993; Ewens y Spielman, 2005). Suponiendo un marcador bialélico, la estratificación se controla comparando las frecuencias alélicas de alelos que se transmite de un padre a un hijo afectado con las frecuencias de alelos que no se transmiten. Existen también muchas extensiones del TDT original que permiten variaciones de múltiples alelos en un locus (Sham y Curtis, 1995), o que permiten realizar TDT sin necesidad se tener los genotipos del hijo afectado y los dos padres, una información no siempre disponible (Sun *et al.*, 1999)

La principal desventaja del TDT es que requiere la recolección de individuos emparentados, con lo que se elimina una de las ventajas de los estudios de asociación (Elston y Spence, 2006).

Para evitar esto se han desarrollado otros métodos que permiten controlar la estratificación de poblaciones sin necesidad de disponer de individuos emparentados.

Si se sospecha que existe estratificación, es posible testarla y controlarla usando marcadores genéticos seleccionados al azar (Pritchard y Rosenberg, 1999; Devlin *et al.*, 2001). En la aproximación llamada *Genomic Control* (Devlin *et al.*, 2001) los autores tratan de estimar y corregir la dispersión, debida a la estratificación, de los estadísticos que miden la asociación de los marcadores con la enfermedad.

Por otra parte Pritchard y Rosenberg han desarrollado métodos que permiten, mediante el uso de marcadores no ligados, hacer inferencias sobre la subestructura poblacional, y usar esta información para testar la asociación genética (Pritchard *et al.*, 2000b; Falush *et al.*, 2003). Existen herramientas como STRUCTURE (Pritchard *et al.*, 2000a) y STRAT (Pritchard *et al.*, 2000b) que permiten detectar estratificación y testar la asociación genética en presencia de ésta.

Otro método alternativo para controlar la estratificación y que se basa en tests de permutación es el llamado *Population Stratification Association Test* (PSAT) (Kimmel *et al.*, 2007). Es un método en el que el test de permutación tiene en cuenta la dependencia entre marcadores y la estructura poblacional y por tanto, a diferencia de un test de permutación estándar, no asume un modelo en el todos los individuos tienen igual probabilidad de tener la enfermedad. PSAT realiza un muestreo a partir de una distribución de probabilidad condicional adecuada (obtenida a partir de cualquier método que estime estructura poblacional, como STRUCTURE), y esos muestreos se utilizan para evaluar la significación estadística de un determinado *score* de asociación (obtenido de los locus más asociados) de forma que se cuenta la fracción de *scores* de asociación permutados que son mayores que el observado.

3. El papel de la bioinformática: del pregenotipado al postgenotipado

Como en muchos otros ámbitos de la biología y la medicina, también en el genotipado la bioinformática juega un papel de importancia creciente, especialmente ahora que los grandes avances tecnológicos han supuesto el abaratamiento de la secuenciación de genomas y la disminución en el coste de genotipado a gran escala (hasta un céntimo de dólar por SNP).

Esta disminución de costes, junto con el acceso a grupos de poblaciones apropiadas, están permitiendo una evaluación exhaustiva de las asociaciones entre variantes genéticas y enfermedad.

Pero a la vez, las tecnologías de genotipado a gran escala están creando un desafío analítico importante; para llegar al descubrimiento y mapeo de los polimorfismos o mutaciones relacionadas con una enfermedad y a la posterior elucidación del mecanismo bioquímico o biofísico que lleva al fenotipo de la enfermedad, es necesaria una combinación de investigación de laboratorio y análisis de datos, y en todo ese proceso la aplicación de métodos bioinformáticos es esencial.

Pregenotipado: selección de SNPs

La secuencia codificante del genoma humano contiene aproximadamente entre 100,000 y 300,000 SNPs codificantes, y otros SNPs adicionales están situados en regiones reguladoras de genes que pueden ser relevantes para el estudio de enfermedades y de salud humana. Los SNPs codificantes y reguladores son de interés particular para los estudios de asociación epidemiológicos. Los SNPs no sinónimos se traducen en cambios de aminoácido en las proteínas que codifican. Los SNPs reguladores pueden afectar la expresión, la especificidad de tejido o la función de proteínas. Parece que tanto unos como otros son relativamente raros comparados con el número total de SNPs en el genoma humano, algo que puede ser consecuencia de la selección en contra de la interrupción de la función causada por estos SNPs.

Uno de los mayores desafíos es elegir los SNPs diana que tienen más posibilidad de afectar al fenotipo, y por tanto contribuir al desarrollo de la enfermedad. Este tipo de variantes suelen priorizarse para su inclusión en los estudios de asociación.

A la hora de seleccionar un conjunto de marcadores optimizado para un estudio de asociación en una región de interés, es necesario identificar todos los SNPs comunes de esa región y seleccionar los tag SNPs basándose en el conocimiento de LD y haplotipos a lo largo de la región. Normalmente lo mas práctico es identificar los genes de esa región y analizar la secuencia codificante con los sitios de *splicing* junto con la región promotora inmediatamente río arriba del sitio de inicio de la transcripción y otros elementos reguladores conocidos, en vez de secuenciar la región entera que puede ser de varias megabases. Se pueden utilizar métodos bioinformáticos para identificar SNP en esas regiones y eliminar los SNPs redundantes con programas de selección de tag SNPs como Tagger (deBakker *et al.*, 2005), o incluso de forma más simple comparando el LD entre SNPs (Zeggini *et al.*, 2005).

También con programas bioinformáticos es posible identificar las variantes genéticas que son más probable que muestren un efecto alélico no neutral, y así forzar su inclusión en las herramientas de selección de tags. En su nivel más simple, la identificación de SNPs potencialmente funcionales

comienza con la identificación de los SNPs localizados en regiones conservadas o en putativos elementos reguladores, es decir, en regiones que están potencialmente conservadas debido a su función. Una vez que un polimorfismo putativamente funcional se ha identificado, el impacto de los diferentes alelos se puede evaluar usando de nuevo la herramienta bioinformática que se usó originalmente para predecir el elemento funcional, como promotores o sitios de *splicing*.

Postgenotipado: análisis e interpretación de los datos

Además de en la preselección de SNPs, la bioinformática juega un papel clave en el análisis e interpretación de los resultados en los estudios de asociación (Campbell y Rudan, 2002).

Partiendo de los resultados en crudo en un estudio de este tipo, hay que llegar a una lista final de genes que se priorizan, apoyándose en una base lógica, para estudiar su posible asociación a la enfermedad. Hay que asegurarse de que se testan los candidatos correctos y aplicar métodos que puedan confirmar el papel biológico de las asociaciones positivas.

En este camino, desde los análisis estadísticos preliminares hasta el testeo de la asociación, hay muchos pasos que pueden facilitarse e incluso no se podrían llevar a cabo sin análisis bioinformáticos. La bioinformática es por tanto clave para tratar problemas como los errores de tipo I, tamaño de muestras (Lohmueller *et al.*, 2003), estratificación de poblaciones (Zang *et al.*, 2007) o errores de genotipado derivados de estudios de asociación de genomas enteros (Plagnol *et al.*, 2007).

Algunos meta-análisis recientes sugieren que la mayoría de las asociaciones que se han encontrado no son correctas, y que esos falsos positivos son probablemente responsables de la cantidad de fracasos en la replicación de asociaciones entre variantes comunes y enfermedades complejas (Lohmueller *et al.*, 2003; Ioannidis *et al.*, 2001).

Estos falsos positivos pueden surgir por el uso inapropiado de P valores por debajo de 0.05 como criterio de significación estadística (Newton-Cheh y Hirschhorn, 2005). Para solucionarlo, el método de corrección de P valores mas común es el de Bonferroni. Sin embargo en muchos escenarios esta corrección es excesiva, ya que asume que todas las variantes que están siendo testadas tiene igual probabilidad *a priori* y no considera la correlación entre variantes genéticas (LD) y entre fenotipos relacionados (como por ejemplo el índice de masa corporal y la circunferencia de la cintura).

El test de permutación proporciona un método empírico que corrige los P valores de forma que retiene la correlación presente en los datos reales controlando mejor la tasa de error. Si por ejemplo

tenemos un fichero con los genotipos, permutando las etiquetas de fenotipos (casos y controles) se mantienen la correlación entre genotipos pero cualquier asociación entre genotipo y fenotipo se perderá. Si se calcula un estadístico con los datos originales y después con los datos permutados miles de veces, el proceso genera una distribución de estadísticos distribuidos según la hipótesis nula de no asociación entre genotipo y fenotipo, permitiendo obtener una significación del experimento. Otros métodos, como el *False Discovery Rate* (Benjamini y Hochberg, 1995, 2000) también se han aplicado para la corrección de P valores, aunque no está claro que este método sea aplicable en una situación donde se espera que la mayoría de los resultados sean falsos positivos (Newton-Cheh y Hirschhorn, 2005).

Otra fuente de falsos positivos que puede corregirse con métodos bioinformáticos es la estratificación de poblaciones. Como se ha comentado anteriormente, la estratificación puede surgir si poblaciones de diferente historia demográfica y diferentes valores medios de un determinado rasgo a estudiar, están mezcladas en el mismo estudio. Como resultado, puede haber una sobrerepresentacion de una subpoblación en el grupo de casos o de controles, y si un alelo genotipado es más común en esa subpoblación, puede aparecer una asociación de ese alelo que podría ser falsa. Se han desarrollado varios métodos bioinformáticos para detectar y controlar esta estratificación mediante el genotipado de marcadores no ligados (Pritchard y Rosenberg, 1999; Rosenberg *et al.*, 2003) y tests de permutación (Kimmel *et al.*, 2007).

Por estas razones, aunque los estudios de asociación son una herramienta poderosa para identificar variantes que influyen en la susceptibilidad a desarrollar enfermedades comunes, la interpretación de estos estudios no es fácil. Aunque el análisis de SNPs individuales puede ser relativamente sencillo, el análisis estadístico de muchos SNPs y sobre todo de los efectos combinados de muchos SNPs puede dar lugar a asociaciones erróneas, y hace necesario disponer de herramientas bioinformáticas que puedan llevar a cabo esos análisis de forma rutinaria.

4. Selección de SNPs

4.1. La importancia de los SNPs funcionales

Las enfermedades genéticas humanas se caracterizan generalmente por un amplio rango de variabilidad fenotípica que se manifiesta a distintas edades, y en distintos grados de severidad o de respuesta al tratamiento. Las causas que subyacen a esta variabilidad están influenciadas por distintos

niveles de modificadores genéticos y medioambientales. Es probable que la mayor parte de las variables genéticas humanas tengan un efecto neutral, pero algunas pueden causar o modificar el fenotipo de la enfermedad. Si la hipótesis de CD/CV es cierta, puede haber un número grande de estas variaciones en las bases de datos de polimorfismos, incluso pueden haber sido también caracterizadas en HapMap. A estas variaciones genéticas se les llama "polimorfismos candidatos".

A medida que el número de SNPs anotados en las bases de datos públicas va creciendo, un objetivo importante en genética humana es la identificación de variantes potencialmente funcionales.

Así como los genes con un putativo papel biológico en la enfermedad se priorizan para su inclusión en los análisis de asociación, los polimorfismos candidatos pueden priorizarse basándose en el efecto predicho en la estructura y función de regiones reguladoras, genes, tránscritos o proteínas.

Un polimorfismo puede afectar casi cualquier proceso biológico. Mucha de la literatura en este campo se centra en la forma más obvia de variación, cambios no sinónimos en regiones codificantes (Ramensky *et al.*, 2002). En las enfermedades genéticas simples estas mutaciones suelen ser mutaciones de aminoácido o mutaciones que producen un codón de terminación, y son fácilmente identificables debido a nuestro conocimiento de las reglas de traducción génica. Es cierto que las alteraciones de la secuencia aminoacídica han explicado un gran número de enfermedades. Sin embargo, en enfermedades complejas está ahora generalmente aceptado que las variaciones que ejercen sus efectos en la susceptibilidad a desarrollar una enfermedad lo hacen a través de mecanismos más sutiles, entre los cuales la alteración de la expresión génica es mayoritaria (Knight, 2005).

Así, los efectos de los polimorfismos de DNA de ninguna manera se limitan a regiones codificantes, las variaciones en regiones reguladoras pueden alterar la secuencia de sitios de unión a factores de transcripción o elementos promotores; variaciones en las zonas UTR del mRNA puede alterar la estabilidad del mRNA; variaciones en regiones reguladoras como potenciadores y silenciadores en exones e intrones pueden alterar la eficacia del *splicing*.

De hecho, para muchos genes las variaciones genómicas que pueden alterar el proceso de *splicing* pueden representar hasta un 50% de todas las mutaciones que conducen a una disfunción génica (Buratti *et al.*, 2001).

El primer paso para identificar SNPs funcionales es determinar la región genómica donde se encuentra. Ésta es la base para elegir herramientas apropiadas ya que son completamente dependientes de la localización de la variación dentro del gen o región reguladora. Parte de este análisis puede hacerse usando visores genómicos como Ensembl (Hubbard *et al.*, 2007) o el buscador

de UCSC (Kuhn *et al.*, 2007). Poner el polimorfismo en su contexto genómico es útil para evaluar las variaciones en términos de localización con respecto al gen (exónicos, intrónicos, UTR, codificantes, región promotora..), o en regiones conservadas. Mooney mostró que en general, las mutaciones asociadas a enfermedad tienden a ocurrir en posiciones que están conservadas (Mooney *et al.*, 2003; Mooney, 2005).

Además de la localización general de variaciones que algunas herramientas bioinformáticas generales pueden ofrecer, hay un contexto mucho más detallado para muchos elementos reguladores conocidos en genes y regiones reguladoras de genes. En términos muy simples, la identificación de SNPs funcionales consiste en identificar SNPs que solapen con esos elementos y evaluar el impacto que sus diferentes alelos tienen en la secuencia original mediante herramientas bioinformáticas.

Sin embargo, muchas veces no es fácil predecir el posible efecto funcional de un SNP, incluso cuando la región en la que está situado esté muy bien caracterizada. La selección de SNPs puede hacerse en estos casos de una forma indirecta, bajo la hipótesis de que la variante estudiada puede estar en LD con la variante funcional. En realidad la mayor parte de los estudios de asociación son una combinación de las dos aproximaciones, donde, si bien domina la aproximación indirecta, se incluyen SNPs con potencial efecto funcional (figura 4).

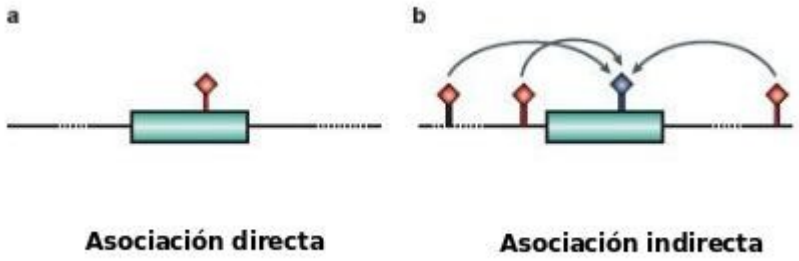


Figura 4. a) *Un caso donde se evalúa de forma directa la asociación de un SNP candidato (en rojo) con una enfermedad. Esta estrategia se utiliza cuando los SNPs se seleccionan para el análisis usando información a priori sobre su posible función, como por ejemplo SNPs que producen un codón de stop, que pueden alterar la función de un gen candidato (rectángulo verde).* ***b)*** *Los SNPs que se van a genotipar (en rojo) se eligen según un modelo de LD de forma que proporcionen información sobre tantos otros SNPs como sea posible. En este caso se evalúa la asociación del SNP de color azul de forma indirecta, ya que está en LD con los otros 3 SNPs. (Figura obtenida de Hirschhorn y Daly, 2005).*

El LD hace más fácil la identificación de genes, pero también hace que sea más difícil e incluso a veces imposible distinguir entre el locus causante de la patología real y sus marcadores correlacionados. Sin embargo proporciona un punto de partida para identificar polimorfismos funcionales.

4.2. SNPs en regiones reguladoras

El primer paso para el estudio de polimorfismos reguladores es determinar si están situados en una región reguladora o en una región codificante. Diferentes pasos de regulación implican distintos elementos como factores de transcripción (TFs), elementos reguladores en cis y otros co-factores (figura 5). Además también implican regiones muy diferentes. Por ejemplo, el **promotor** es la región reguladora más importante que controla y regula el primer paso de la expresión génica, la transcripción del mRNA. La señal de ***splicing*** está en los llamados sitios de *splicing* que bordean los exones, y está fuertemente regulada por secuencias exónicas e intrónicas que pueden actuar como potenciadores o silenciadores de *splicing*. La regulación transcripcional también depende de manera importante de la **estructura** de la cromatina (Wasserman y Sandelin, 2004).

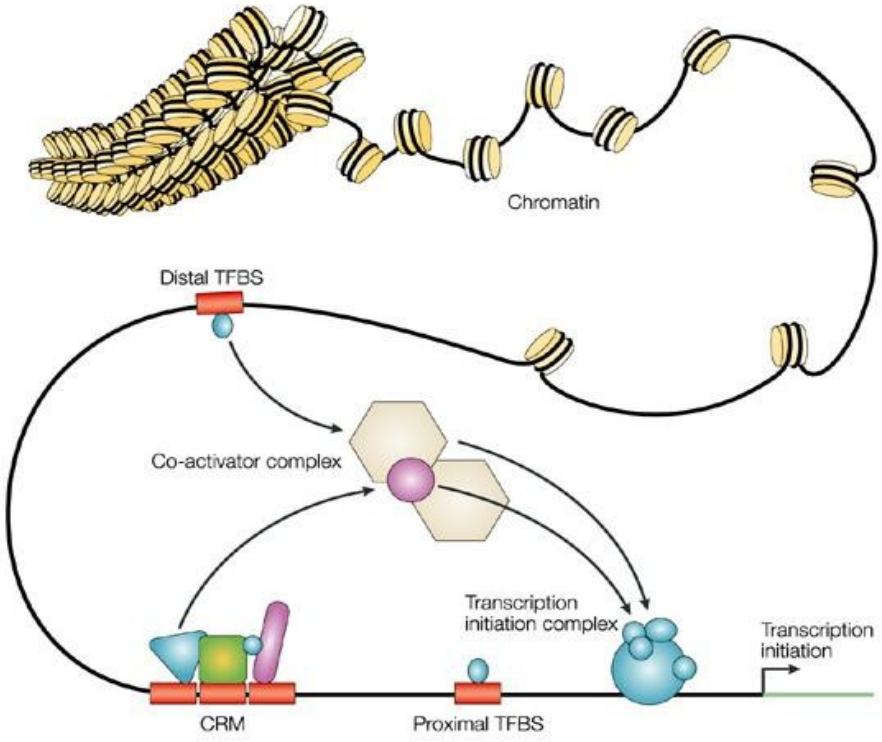


Figura 5. *Los TFs se unen a sitios específicos (sitios de unión a factores de transcripción, TFBSs) que son proximales o distales con respecto al sitio de inicio de la transcripción. Los TFs pueden agruparse y operar en módulos reguladores en cis para conseguir propiedades reguladoras específicas. Las interacciones entre TFs unidos al DNA y otros co-factores estabiliza la maquinaria de inicio de la transcripción para permitir la expresión génica. La regulación conferida por la unión específica de secuencia de los TFs es altamente dependiente de la estructura de la cromatina. (Figura obtenida de Wasserman y Sandelin, 2004).*

Para organismos como humano o levadura, cuyas anotaciones genómicas son relativamente completas, los servidores web de genomas son herramientas muy útiles para identificar estructuras génicas y otras anotaciones relacionadas (Kuhn *et al.*, 2007; Hubbard *et al.*, 2007). Estos servidores incluyen tanto genes anotados manualmente como genes predichos computacionalmente. Muchos otros recursos, incluyendo bases de datos de promotores y métodos computacionales para la predicción de promotores, también están disponibles para poder caracterizar promotores de una forma bastante precisa.

4.2.1. Promotores

Un promotor se define normalmente como una región de DNA, cercana al sitio de inicio de la transcripción (TSS), que es necesaria para controlar y regular el inicio de la transcripción del gen que le sucede río abajo. En humanos, para que la transcripción se inicie de forma eficiente, es necesario el ensamblado en el promotor de un complejo multiproteico que contiene a la DNA polimerasa II y seis factores de transcripción (TFs) generales, IIA, IIB, IID, IIE, IIF y IIH (Lagrange *et al.*, 1998). Este ensamblado requiere la presencia en el promotor de un número de elementos con secuencias consenso, como el elemento de reconocimiento para el factor TFIIB (BRE), el elemento promotor río abajo (DPE) (Burke y Kadonaga, 1996), el iniciador (Inr) (Smale y Baltimore, 1989) y especialmente la caja TATA (Smale y Kadonaga, 2003). Los promotores que contienen la caja TATA se descubrieron primero en organismos bacterianos, y se pensó que esta caja era el elemento promotor universal. Más tarde se descubrieron promotores humanos sin caja TATA, y su porcentaje ha ido creciendo desde entonces, desde un 22% (Bucher, 1990) hasta un 78% (Gershenzon y Ioshikhes, 2005).

Hoy en día se considera que la arquitectura del promotor con la clásica caja TATA representa una minoría de los promotores en mamíferos, siendo esta clase de promotores comúnmente asociada con genes específicos de tejidos, y la abundancia de islas CpG, que es la característica dominante de las secuencias promotoras en humanos, junto con otras características adicionales, ha cambiado el objetivo de los algoritmos para la predicción computacional de promotores humanos.

Las características importantes para los programas de predicción de promotores incluyen el contenido en GC, el ratio CpG, la densidad de sitios de unión a factores de transcripción, la composición de secuencias cortas y los elementos promotores núcleo como la caja TATA, DPE o Inr. Algunos de estos programas son por ejemplo, PromoterScan2 (Prestridge, 1995), Eponine (Down y

Hubbard, 2002), y PromoterInspector (Scherf *et al.*, 2000).

En general, aunque los promotores proximales pueden no contener toda la información necesaria para controlar de forma precisa la transcripción de genes en espacio y tiempo, el análisis de los promotores solos puede generar modelos significativos de redes reguladoras transcripcionales.

Sitios de unión a factores de transcripción (TFBSs)

Los objetivos en la predicción de promotores y la identificación de TFBSs no son exactamente los mismos. Mientras que la predicción de promotores trata de localizar el TSS y sus regiones reguladoras, el objetivo de los métodos computacionales para el modelado y predicción de TFBSs es entender interacciones cis-trans para la regulación de la transcripción.

La mayoría de los TFBSs son secuencias cortas de 6-20 bases localizadas en regiones no codificantes de gen, casi siempre en la zona 5' aunque a veces en 3' o incluso en intrones.

Sin embargo sólo entre 4 y 6 bases dentro de cada TFBS están completamente conservadas, y el resto son altamente variables. Como resultado, los TFBSs normalmente se modelan utilizando matrices de pesos específicas de posiciones (PWMs), basadas en alineamientos de sitios conocidos, determinados experimentalmente. Dichas matrices esencialmente resumen las frecuencias relativas de cada nucleótido en cada una de las posiciones del TFBS.

La estructura de la matriz nos permite asignarle una puntuación cuantitativa a cualquier secuencia para identificar sitios de unión potenciales (Wasserman y Sandelin, 2004).

Para reducir el gran número de falsos positivos que se generan debido a la degeneración de los TFBSs, normalmente se impone un criterio de conservación a la región reguladora conocido como *phylogenetic footprinting*, y que se se refiere a la identificación de regiones funcionales mediante la comparación de secuencias genómicas ortólogas entre especies (Fickett y Wasserman, 2000; Zhang y Gerstein, 2003).

Con la disponibilidad de un mayor número de genomas secuenciados, los análisis comparativos de regiones no codificantes han llegado a ser una aproximación importante para detectar promotores o regiones reguladoras en general (Bejerano *et al.*, 2004; Siepel *et al.*, 2005) y este procedimiento mejora significativamente el poder de la predicción de TFBSs, como se demuestra en el ejemplo descrito en detalle en Lenhard *et al.* (Lenhard *et al.*, 2003).

4.2.2. Splicing

En organismos eucariotas, muchos genes están interrumpidos por secuencias no codificantes llamadas intrones. Estos intrones se transcriben en el mRNA pero, antes de la traducción son eliminados mediante un proceso conocido como *splicing*. Un gen con varios exones puede procesarse de varias maneras (incluyendo distintos exones), proceso conocido como *splicing* alternativo.

La regulación génica a través del *splicing* alternativo es más versátil que la regulación a través de la actividad promotora. Los cambios en la actividad promotora alteran predominantemente los niveles de expresión del mRNA. En cambio, cambios en el *splicing* alternativo pueden modular los niveles de expresión génica sometiendo al mRNA a una degradación mediada por mutaciones terminadoras (NMD) (Maquat, 2004) y alterando la estructura del producto génico insertando o delecionando partes proteicas. Los efectos causados por variaciones en el *splicing* alternativo van desde una pérdida completa de la función a efectos sutiles que son difíciles de detectar.

La regulación del *splicing* está mediada por el spliceosoma, un macro-complejo compuesto de ribonucleoproteína nuclear pequeña (RNPnp) y de la familia de proteínas ricas en serina/arginina (SR). En su nivel más básico, el *splicing* de premRNA implica la eliminación precisa de los intrones para formar el mRNA maduro, con una pauta de lectura intacta. Un *splicing* correcto implica el reconocimiento de los exones y el corte y empalme precisos en las fronteras exónicas designadas por los dinucleótidos invariables GT y AG, conocidos como los sitios donador y aceptor.

Mutaciones en el sitio donador normalmente causan la pérdida (*skipping*) de su exón asociado (Carmel *et al.*, 2004) y a veces producen eventos adicionales como la inclusión completa del intrón (Zhang *et al.*, 2004) o la activación de sitios de *splicing* alternativos conocidos como sitios de *splicing* crípticos. De hecho, la activación del sitio de *splicing* críptico del gen de la β-globina fue uno de los primeros defectos de *splicing* con relevancia médica que se han descrito (Wieringa *et al.*, 1983; Treisman *et al.*, 1983).

Cada dinucleótido está flanqueado por una secuencia más larga y menos conservada. El sitio de ramificación y la región de polipirimidinas cercanos al extremos 3' del intrón también son críticos para el *splicing* (figura 6). Sitios menores de *splicing* como por ejemplo intrones “AU-AC”, aunque son menos del 0.1%, también existen (Burset *et al.*, 2000).

Lo que se tiene claro hoy en día es que los dos elementos dinucleótidos de *splicing* consenso, aunque necesarios, no son suficientes para definir los límites intrón-exón.

Para aumentar la fidelidad total de la reacción de *splicing*, existen otras secuencias adicionales en exones e intrones. Esos elementos de secuencia que actúan en cis pueden actuar aumentando o disminuyendo el reconocimiento y se llaman respectivamente potenciadores y silenciadores de *splicing* exónicos (ESE, ESS) e intrónicos (ISE, ISS).

Los potenciadores y silenciadores están implicados en *splicing* constitutivo y alternativo, y en la mayoría de los casos no tienen una secuencia consenso bien definida. Además, esos elementos no están siempre definidos inequívocamente y sus funciones se pueden solapar.

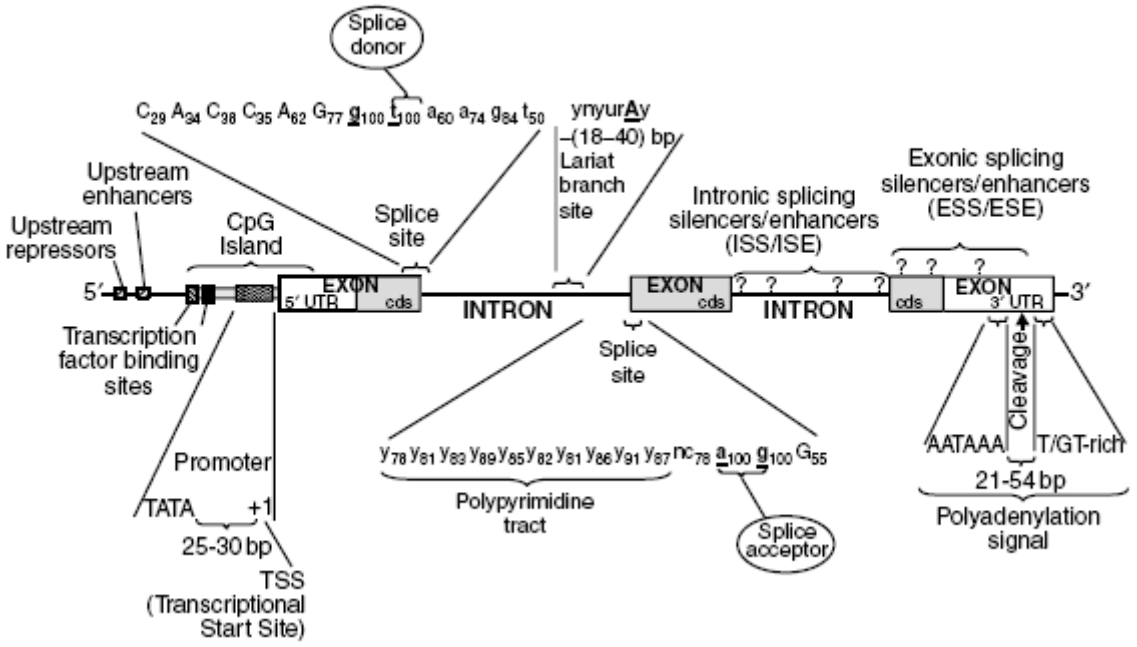


***Figura 6**. La figura muestra algunas de las regiones reguladoras clave que controlan la transcripción, el splicing y el procesamiento post-transcripcional de genes y tránscritos. Los polimorfismos en esas regiones son potenciales SNPs con efecto funcional. (Figura obtenida de Barnes y Gray, 2003).*

Los ESEs han sido sujetos a muchos estudios y la mayoría, aunque no todos, se sabe que son reconocidos por miembros de la familia de proteínas SR. En particular, los potenciadores exónicos ricos en A/C, a través de SELEX funcional, se ha visto que tienen un papel importante en el reconocimiento de exones. Las proteínas SR se unen a ESEs a través de dominios de unión al RNA y promueven el *splicing* reclutando componentes del spliceosoma a través de interacciones proteína-proteína por medio de dominios ricos en arginina/serina.

Hay dos programas de predicción de ESEs que están disponibles actualmente, ESEFinder (Cartegni *et al.*, 2003) y RESCUE-ESE (Fairbrother *et al.*, 2002).

El primero es una herramienta web que proporciona matrices de pesos de secuencias para puntuar un subconjunto de motivos ESE candidatos correspondientes a los motivos consenso funcionales de 4 proteínas SR, incluyendo SF2/ASF, SC35, SRp40 y SRp55 que fueron identificadas a través de un método SELEX funcional (Tuerk y Gold, 1990). El valor predictivo de esas matrices se ejemplifica por el hecho de que motivos con puntuaciones altas están enriquecidos en exones, agrupados en regiones que contienen ESEs naturales, y por correlaciones entre puntuaciones de motivos y fenotipos de *skipping* de exones en varios genes (Cartegni *et al.*, 2003).

RESCUE-ESE predice motivos con secuencias ESE basadas en el análisis estadístico de diferencias en frecuencias de hexámeros entre exones e intrones y entre exones con sitios de *splice* fuertes y débiles. Aunque está menos corroborado que ESEfinder, ha identificado correctamente secuencias que actúan como ESEs y representantes de 10 motivos predichos que se ha mostrado que tiene actividad potenciadora en minigenes indicadores (Fairbrother *et al.*, 2002).

Como ejemplo, recientemente se ha descrito una mutación en un ESE del exón 3 del gen MLH1 que produce cáncer colorectal hereditario no-polipósico (McVety, 2006). El motivo ESE identificado como el responsable no es reconocido por ESEFinder y si por RESCUE-ESE, aunque esta herramienta también predice otros dos que podrían ser falsos positivos.

Los factores que se unen a ESSs no se han caracterizado con el mismo detalle, sin embargo, se ha visto que algunas ribonucleoproteínas nuclear heterogéneas (RNPnh) podrían están implicadas en interacciones con estos elementos (Baralle y Baralle, 2005).

Dos grupos han usado métodos computacionales para predecir ESSs (Sironi *et al.*, 2004; Zhang y Chasin, 2004). Los dos métodos asumen que los ESSs están enriquecidos en pseudo-exones en comparación con exones reales. Sironi y colaboradores predijeron 3 motivos ESS, uno de ellos, con secuencia similar al sitio de unión para RNPnh H, se confirmó experimentalmente. Zhang y colaboradores predijeron 974 putativos ESSs 8-mer usando 2 criterios, enriquecimiento en pseudo-exones relativo a exones no codificantes y enriquecimiento en 5'UTR sin intrones relativo a exones no codificantes. Los 974 ESSs se agruparon en 69 familias, cuyas secuencias consenso generalmente no coinciden de forma exacta con motivos ESS conocidos.

Posteriormente el grupo de Burge (Wang *et al.*, 2004) realizó un cribado sistemático para ESSs. Este cribado identificó 141 decámeros ESSs. Esos decámeros pudieron agruparse, de acuerdo con su similitud de secuencia, en siete grupos que dieron lugar a 7 putativos motivos ESS, cuya secuencia se parece a los sitios de unión conocidos para RNPnh H y A1. Esos decámeros se analizaron buscando un motivo núcleo consenso y se encontró que tenían un enriquecimiento significativo de 103

hexámeros, el Fas-hex-3 set, que podrían ser los motivos núcleo consenso de los ESSs (Wang *et al.*, 2004).

Todavía se sabe menos aún de los mecanismos por los cuales funcionan los ISEs e ISSs, aunque se han descrito mutaciones intrónicas que actúan como potenciadoras (Ishii *et al.*, 2002) y silenciadoras (D'Souza y Schellenberg, 2000), y se ha descrito que secuencias repetitivas GT podrían actuar como ISEs en la regulación de la expresión del gen NCX1 (Gabellini, 2001). Sin embargo, debido a que los mecanismos no están claros, el único modo de valorar si hay mutaciones afectando al *splicing* es testándolas experimentalmente.

4.2.3. Estructura del DNA

La transcripción está modulada por los factores de transcripción y elementos reguladores en cis, pero también pero hay varios estudios, tanto experimentales como computacionales, que muestran que las regiones promotoras poseen un número de características dependientes de secuencia que los hace distintos del resto del genoma, como su flexibilidad, curvatura o estabilidad (Kanhere y Bansal, 2005) y que estas características tienen una gran influencia en el proceso de transcripción (Wasserman y Sandelin, 2004).

Se ha sugerido que los tríplex de DNA (Pauling y Corey, 1953; Felsenfeld *et al.*, 1957) podrían ser regiones reguladoras que pueden controlar la expresión génica (Goñi *et al.*, 2004). Las secuencias capaces de formar triple hélices (*triplex-forming oligonucleotide target sequences*, TTSs) son secuencias de más de 10 polipirimidinas o polipurinas, cuya presencia es mucho más abundante de lo esperado a partir de modelos aleatorios simples (Goñi *et al.*, 2004). Se ha visto que la mayor concentración de TTSs se encuentra en regiones reguladoras, especialmente en zonas promotoras, lo que sugiere una tremenda potencialidad de estas secuencias en el control de la expresión génica (Goñi *et al.*, 2004). Aunque el mecanismo por el cual actúan es muy especulativo, se cree que tiene que ver con la flexibilidad del DNA.

La flexibilidad es la facilidad con la que la molécula se puede curvar en cualquier dirección, y esta flexibilidad, que depende de su secuencia (Tsai *et al.*, 2002), puede permitir interacciones de proteínas unidas al DNA en sitios diferentes (Tsai *et al.*, 2002), o evitar impedimentos estéricos (Buckland, 2006). Algunos análisis computacionales sugieren que la curvatura intrínseca del DNA puede ser un criterio importante para el reconocimiento de la caja TATA (Nishikawa *et al.*, 2003). Además muchos

promotores sin caja TATA también contienen frecuentemente una estructura de DNA curvada, lo que indica que esta curvatura puede jugar un papel importante con independencia del tipo de promotor (Nishikawa *et al*., 2003). Pedersen y colaboradores estudiaron genomas procariotas y encontraron una tendencia a que el DNA promotor esté más curvado, menos flexible y menos estable que el DNA en regiones codificantes y el DNA intergénico sin promotores (Pedersen *et al.*, 2000).

La formación de tríplex en las zonas promotoras podría afectar al grado de flexibilidad y a la curvatura del DNA en esas zonas y por tanto favorecer o perjudicar la interacción entre factores de transcripción o entre factores de transcripción y sus sitios de unión al DNA (figura 7).

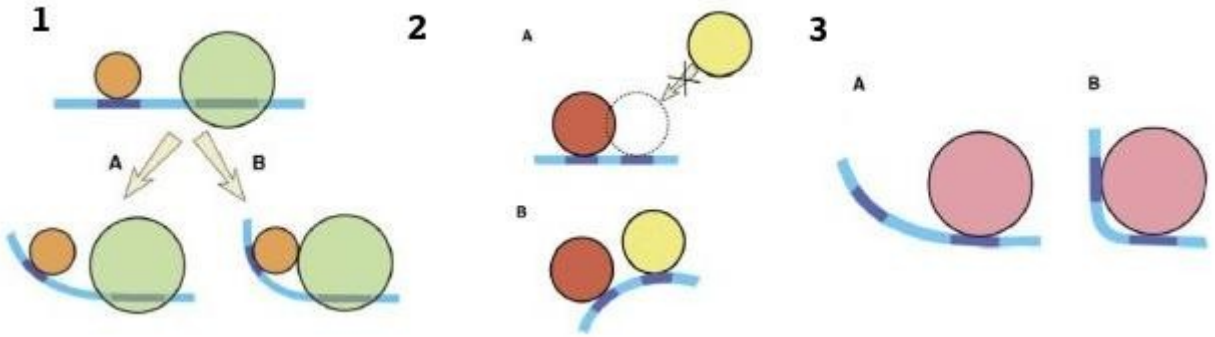


Figura 7*. 1) La flexibilidad del DNA permite la interacción entre TFs, cuanto más flexible es el DNA, con mayor facilidad ocurre la interacción.* ***2)*** *La curvatura del DNA evita el impedimento estérico, en 2A las proteínas no se pueden unir al DNA al mismo tiempo, en 2B la curvatura permite más espacio para la unión simultánea de las dos proteínas.* ***3)*** *Para permitir la unión de una proteína en dos sitios de unión del DNA a la vez, la secuencia de DNA entre los dos sitios de unión debe ser curva o flexible como en 3B (Figuras obtenidas de Buckland, 2006).*

4.3. SNPs codificantes no sinónimos (nsSNPs)

Los SNPs que se encuentran en las regiones codificantes de los genes son a menudo SNPs no sinónimos, es decir, que cambian un aminoácido en la secuencia proteica en la que se encuentran. Estos SNPs pueden ser neutrales, donde la proteína mutada no se distingue funcionalmente de la proteína normal, o no neutrales, donde la proteína mutada puede tener su función alterada respecto a la normal. Estos SNPs, junto con los SNPs situados en regiones reguladoras, son los que probablemente tengan el mayor impacto en el fenotipo (Ramensky *et al.*, 2002).

Hay muchas formas por las que un nsSNP puede afectar a la función de la proteína. Lo más probable es una pérdida parcial o completa de la función. Algo menos probable es una mutación que

produce una ganancia de función, como la observada en la activación del oncogen RAS (Quilliam *et al.*, 1995)

Existen varias aproximaciones para la predicción de la función de nsSNPs, que incluyen estudios de propiedades basadas en secuencia, propiedades estructurales y propiedades derivadas de alineamientos de secuencias o filogenias (Mooney, 2005). Entre las distintas aproximaciones empleadas se encuentran métodos de reglas empíricas (Ng y Henikoff, 2001), árboles de decisión (Dobson *et al.*, 2006, Krishnan y Westhead, 2003), máquinas de soporte de vectores (Bao *et al.*, 2005), redes neuronales (Ferrer-Costa *et al.*, 2002, 2004, 2005), redes bayesianas (Cai *et al.*, 2004) o métodos de estima de presión selectiva (Arbiza *et al.*, 2006). Aunque cada método es distinto, en general, para clasificar un mutación como patológica, casi todos entrenan un predictor con un conjunto de entrenamiento formado por mutaciones patológicas conocidas, ya sea a través de estudios mutagénicos (Krishnan y Westhead, 2003; Cai *et al.*, 2004; Ng y Henikoff, 2001), a través de las anotadas en bases de datos (Bao y Cui, 2005) o usando pseudo-mutaciones entre proteínas ortólogas en especies cercanas evolutivamente (Ferrer-Costa *et al.*, 2002, 2004, 2005).

Existen hoy en día numerosas herramientas web disponibles para la predicción de mutaciones no sinónimas patológicas como PolyPhen (Ramensky *et al.*, 2002), SIFT (Ng y Henikoff, 2003), nsSNPAnalyzer (Bao *et al.*, 2005), LS-SNP (Karchin *et al.*, 2005) y Pmut (Ferrer-Costa *et al.*, 2002, 2004, 2005).

Además existen numerosas bases de datos de nsSNPs que incluyen sus propios métodos de análisis como SNP Function Portal (Wang *et al.*, 2006), TopoSNP (Stitziel *et al.*, 2004), PolyDoms (Jegga *et al.*, 2007) o SNPeffect (Reumers *et al.*, 2005, 2006). Ésta última es una base de datos de SNPs codificantes no sinónimos potencialmente funcionales, que utiliza distintas herramientas bioinformáticas basadas en secuencias o estructuras como FoldX (Fernandez-Escamilla *et al.*, 2004), Tango (Schymkowitz *et al.*, 2005), AmyScan (López de la Paz y Serrano, 2004), Psort II (Nakai y Horton, 1999), O-GlycBase (Gupta *et al.*, 1998) o Phosphobase (Kreegipuu *et al.*, 1999), para predecir el efecto que esos SNPs tienen en la estabilidad, dinámica, procesamiento postraducional o localización celular de las proteínas que los contienen.

5. Análisis de datos procedentes de estudios de asociación.

5.1. Análisis preliminar de los datos

El análisis adecuado de los datos genéticos obtenidos en estudios de asociación requiere una observación de las propiedades básicas de los datos, seguido de análisis más especializados.

La calidad de los datos es muy importante en cualquier tipo de análisis, y particularmente en análisis de este tipo la comprobación del HWE puede ser muy útil. Las desviaciones del HWE pueden ser debidas a la estratificación de poblaciones, incluso puede ser un síntoma de asociación a la enfermedad (Balding, 2006). También pueden aparecer desviaciones aparentes en presencia de polimorfismos de deleción o por errores de genotipado, como una mutación en el cebador usado en la PCR o por la tendencia a llamar heterocigotos a los homocigotos (Balding, 2006).

Normalmente el HWE se utiliza como parámetro de calidad para descartar polimorfismos que se desvían del HWE en la población control con un nivel de significación de $\alpha=10^{-3}$ o 10^{-4}. Sin embargo, antes de descartar esos loci, habría que considerar si esa desviación se debe a deleciones o duplicaciones que podrían ser importantes en el desarrollo de la enfermedad.

Otro factor relacionado con la calidad de los datos es el tratamiento de datos incompletos. Para el manejo de datos incompletos existen numerosos métodos. El más simple es el llamado análisis de casos completos, en el que se eliminan individuos que no tiene datos en alguno de los SNPs. Este método puede ser muy ineficiente al disminuir el tamaño de muestra por la eliminación de datos de un SNP de interés pertenecientes a un caso que contiene ausencias en algún otro SNP. Una solución a este problema consiste en la sustitución de los genotipos incompletos con valores predichos a partir de los genotipos observados en SNPs vecinos, los llamados métodos de imputación (Dai *et al.*, 2006; Souverein *et al.*, 2006; Croiseau *et al.*, 2007), entre los que se incluyen asignar la media al valor ausente, predecir el valor ausente mediante modelos de regresión, métodos de máxima verosimilitud e imputaciones múltiples. Sin embargo los métodos de imputación se basan en que la falta del dato es independiente tanto del fenotipo como del genotipo real, algo que no siempre es cierto ya que variantes heterocigotas pueden estar ausentes más a menudo que las homocigotas, e incluso pude haber diferentes tasas de genotipos incompletos entre casos y controles si éstos se obtienen de forma

diferente (Clayton *et al.*, 2005).

La estadística relacionada con el genotipado masivo está en pleno desarrollo como respuesta a las necesidades actuales de los investigadores. Existen varios paquetes estadísticos que calculan estadísticos básicos y realizan análisis más sofisticados que pueden ser utilizados para el análisis preliminar de los datos (Excoffier y Heckel, 2006).

5.2. Métodos de análisis de asociación

Con las nuevas tecnologías de alto rendimiento y con los recursos genómicos con los que se dispone hoy en día, es posible la identificación de genes y polimorfismos genéticos implicados en el desarrollo de enfermedades (Glazzier *et al.*, 2002).

En el análisis de estos datos genéticos, una aproximación sencilla que se usa habitualmente es la evaluación de SNPs individuales. En esta estrategia cada SNP se evalúa con algún procedimiento adecuado, como un test de Chi cuadrado, y de esta forma se identifican los SNPs con una asociación significativa a la enfermedad. Los métodos que se centran en la asociación de alelos de SNPs individuales con la enfermedad son útiles para estudiar enfermedades monogénicas.

Sin embargo, en enfermedades complejas, es posible que muchos SNPs participen en el desarrollo de la enfermedad, aunque la contribución de cada SNP individual sea pequeña o incluso ausente ya que es posible que ciertos loci estén contribuyendo al desarrollo de la enfermedad sólo por sus interacciones con otros genes (epistasis). En estos casos las aproximaciones que se centran en marcadores individuales a menudo no son capaces de encontrar una asociación significativa (Joo *et al.*, 2005).

En los últimos años se han desarrollado métodos que incorporan la naturaleza multigénica de enfermedades complejas a la hora de detectar SNPs con asociación a la enfermedad. Estas aproximaciones se conocen como métodos *multi-marker* (Hoh y Ott, 2003) y han sido utilizadas para el análisis de datos de estudios de asociación genéticos.

Entre las distintas aproximaciones *multi-marker* se encuentran métodos estadísticos tradicionales, como regresión logística (Nagelkerke *et al.*, 2005) o redes neuronales (Tomita *et al.*, 2004), y métodos no-paramétricos, como métodos de *random forests*, métodos combinatoriales o las llamadas aproximaciones *Two Steps*, más exitosas a la hora de manejar números grandes de predictores e identificar interacciones gen-gen (Bureau *et al.*, 2005).

5.2.1. Modelos estadísticos clásicos

Regresión logística

El modelo de regresión logística es un modelo estadístico para estudiar la dependencia de un fenotipo binomial (casos y controles) en un conjunto de factores de riesgo. La probabilidad para una de las dos clases de fenotipo se expresa en forma de su logit (log(p/(1-p)), que se predice por la combinación lineal de los factores de riesgo. Para genotipos, esta combinación lineal es la suma ponderada de los genotipos codificados como 0, 1 o 2 en cada marcador. Los pesos se determinan de forma que la suma discrimina de la mejor forma posible entre los casos y controles (Hoh y Ott, 2003).

El principal problema de este modelo surge cuando el número de marcadores es mayor que el de individuos (problema de dimensionalidad) algo que ocurre generalmente con datos genéticos y que tradicionalmente suele sortearse con el análisis de un solo marcador cada vez. Además el modelo de regresión impone relaciones fijas entre genotipos y fenotipo (casos *versus* controles), una situación que puede no ser realista (Hoh y Ott, 2003).

Redes neuronales

En respuesta a la limitación del modelo de regresión logística, en 2003 Ritchie y colaboradores desarrollaron un método, llamado *Genetic Programming Optimized Neural Network* (GPNN), basado en redes neuronales y optimizado para mejorar la selección de predictores asociados a enfermedad (Ritchie *et al.*, 2003a). Las redes neuronales son un tipo de método de reconocimiento de patrones desarrollado en los años 40, que ha sido utilizado entre muchas otras cosas para determinar predictores genéticos y/o ambientales relacionados con enfermedad en estudios genéticos (Tomita *et al.*, 2004). En concreto, una red neuronal de tipo perceptrón se compone generalmente de una capa de entrada, una o varias capas intermedias y una capa de salida (figura 8). Cada capa, formada por nodos, esta conectada con la siguiente capa y a cada conexión se le asigna un peso. Cada nodo tiene asociada una función matemática denominada función de transferencia, que genera la señal de salida del nodo a partir de las señales de entrada. La reorganización de las conexiones (la estimación de los parámetros de la función de transferencia) se modelan mediante el ajuste de los pesos durante la fase de aprendizaje.

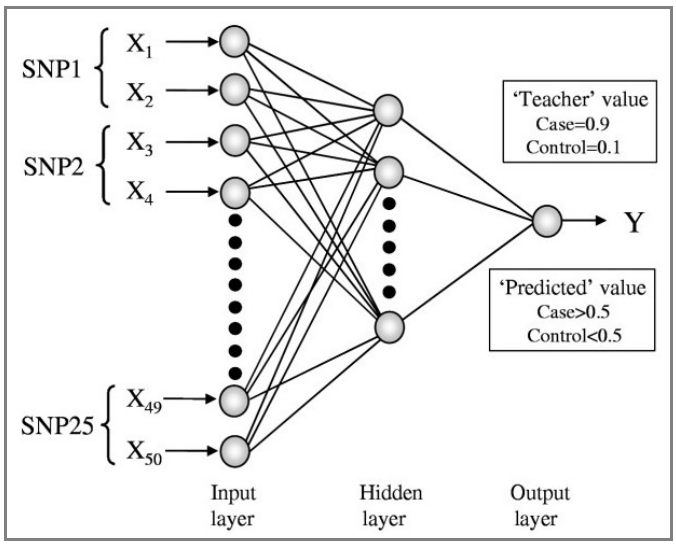


Figura 8. *Modelo de red neuronal propuesto por Tomita y colaboradores. Este modelos se aplicó para analizar la relación entre asma infantil y 25 SNPs en 17 genes candidatos. (Figura obtenida de Tomita et al., 2004).*

Una de sus mayores ventajas es su habilidad para aprender la relación entre variables independientes y una variable resultado en un conjunto de datos, y a partir de ahí hacer predicciones en datos donde la variable resultado es desconocida. Una desventaja es que los parámetros de entrada y la arquitectura de la red debe ser especificada previamente y no hay una regla empírica para generarla por lo que muchas veces hay que realizar procesos de ensayo y error.

Para evitarlo, el algoritmo del GPNN optimiza no solo los pesos sino también los parámetros de entrada, que se seleccionan a partir de un número grande de predictores, y optimiza también la conectividad de la red, el número de capas ocultas y el número de nodos de esas capas para generar la arquitectura óptima de la red neuronal para un determinado conjunto de datos (Ritchie *et al.*, 2003a).

El método de GPNN no está sujeto al problema de la dimensionalidad ya que sólo utiliza una selección al azar de predictores para construir unos modelos iniciales que evolucionan durante el proceso hasta el modelo de mejor estructura, que es evaluado por validación cruzada (Heidema *et al.*, 2006)

El GPNN se aplicó en un estudio de casos y controles en la enfermedad de Parkinson (Motsinger *et al.*, 2006), y mediante datos simulados se mostró que el método tiene una eficacia elevada en la detección de interacciones gen-gen y gen-medio ambiente cuando se aplica a modelos de interacción de 2 y 3 marcadores en tamaños de muestras moderadas (Motsinger *et al.*, 2006)

5.2.2. Métodos no-paramétricos

Random Forests (RF)

Un RF es una colección de árboles de clasificación que crecen desde el nodo raíz por medio de muestreos *bootstrap* de los datos observados, utilizando un subconjunto de predictores al azar para definir el mejor corte en cada nodo. En datos de casos y controles, los datos observados son los individuos y los predictores son los marcadores. Las observaciones que se dejan fuera de las muestras *bootstrap* se utilizan para estimar el error de predicción, de forma que los marcadores de los individuos que se dejan fuera determinan a que nodo o clase se asigna ese individuo en un determinado árbol.

La importancia de un marcador en presencia del resto de marcadores se mide por un índice de importancia I_M.

La aplicación de RF de árboles de clasificación se ha utilizado en el contexto de estudios de asociación de casos y controles, tanto para la clasificación de individuos (Schwender *et al.*, 2004) como para la identificación de SNPs de susceptibilidad en asma (Bureau *et al.*, 2005). En este último estudio, los autores extendieron la noción del índice de importancia para evaluar el valor predictivo de pares de SNPs y analizar así interacciones entre marcadores. Ellos sugieren que cuando muchos marcadores contribuyen al riesgo a la enfermedad, el medir la importancia de pares de SNPs puede ser una aproximación más poderosa que el medir la importancia de cada SNP de forma individual.

Debido a que los I_M para pares de SNPs proporcionan información sobre las interacciones entre SNPs (Bureau *et al.*, 2005), con el método de RF se pueden detectar marcadores que por si solos tengan efectos débiles pero que interaccionen significativamente con otros marcadores (Heidema *et al.*, 2006). Lunetta y colaboradores compararon el método de RF con el test de Fisher en un conjunto de datos simulados, y comprobaron que en presencia de interacciones entre marcadores, la aproximación de RF tenía un mejor rendimiento a la hora de seleccionar SNPs de riesgo (Lunetta *et al.*, 2004).

Métodos combinatoriales

Los métodos combinatoriales buscan sobre todas las posibles combinaciones de factores para encontrar las combinaciones que explican mejor la variable resultado. Un ejemplo de este tipo de métodos que se ha aplicado a estudio de datos genéticos es el *Combinatorial Partitioning Method*

(CPM). Este método (Nelson *et al.*, 2001) originalmente desarrollado para estudiar el efecto de fenotipos cuantitativos, también se ha utilizado para la detección de interacciones gen-gen.

El CPM determina las combinaciones de loci que predicen variación en los niveles cuantitativos del fenotipo o rasgo de estudio, y al mismo tiempo define grupos de genotipos con medias fenotípicas similares. Estos grupos de genotipos se denominan particiones. A las combinaciones de 2 o más particiones se denomina conjunto de particiones genotípicas. El método consta de varios pasos:

i) Se seleccionan combinaciones de loci de entre todas las combinaciones posibles. Por ejemplo, si hay 10 loci y se consideran combinaciones de 2 loci, el número de combinaciones a analizar serán $10(10-1)/2 = 45$.

ii) Para cada una de estas combinaciones se crean las particiones. Por ejemplo, para un par de loci bialélicos habrá 9 posibles particiones (AABB, AABb, Aabb, AaBB, AaBb, Aabb, aaBB, aaBb y aabb).

iii) Esas particiones se combinan para formar conjuntos de particiones genotípicas. Un conjunto puede por ejemplo consistir en 2 particiones, una con los genotipos AABB, AABb y Aabb, y la otra con el resto de genotipos.

iv) De todos los conjuntos se seleccionan aquellos que tienen un mayor efecto en el fenotipo mediante un análisis de la varianza. Estos conjuntos se evalúan por validación cruzada y los conjuntos con mayor poder de predicción se utilizan para hacer inferencias sobre las relaciones genotipo-fenotipo.

Si por ejemplo hay un rasgo cuantitativo influenciado por un locus bialélico donde el alelo *A* es dominante sobre el alelo *a*, el objetivo es i) identificar el alelo *A* como el que predice la variabilidad del rasgo y ii) agrupar genotipos que son fenotípicamente similares en particiones genotípicas {*AA, Aa*} y {aa}, enfatizando similitudes entre genotipos dentro de la partición así como diferencias entre particiones (Nelson *et al.*, 2001).

Debido al coste computacional que supone testar todas las posibles combinaciones, este método es prohibitivo cuando se quieren analizar interacciones que implican más de dos loci. (Moore *et al.*, 2002). Para solventar esto, Culverhouse y colaboradores (Culverhouse *et al.*, 2004) desarrollaron posteriormente una extensión del CPM denominada *Restricted Partition Method* (RPM). Al contrario que los métodos precedentes, el RPM no evalúa todas las posibles combinaciones, sino que descarta aquellas con una varianza grande (si en el grupo de genotipos de esa partición hay diferencias grandes entre los valores fenotípicos medios). En contraste con la aproximación exhaustiva del CPM, el

algoritmo del RPM trata de buscar las particiones más razonables, haciendo un balance entre la maximización de la variación entre grupos con la minimización de la variación intra-grupo (Culverhouse *et al.*, 2004).

Aunque el RPM se desarrolló originalmente para datos cuantitativos, también se ha aplicado de forma satisfactoria a datos de casos y controles con datos simulados (Culverhouse, 2007).

Otra modificación o extensión del CPM es el método *Multifactor Dimensionality Reduction* (MDR), desarrollado para analizar efectos genéticos y/o ambientales en una variable binaria como en casos y controles, en vez de en fenotipos cuantitativos. El método trata de identificar combinaciones de genotipos y factores medioambientales discretos asociadas a un alto riesgo a desarrollar la enfermedad, así como combinaciones asociadas a riesgos bajos. Para ello el método define una sola variable que incorpora información de varios loci y/o factores medioambientales y que puede asociarse a combinaciones de alto y bajo riesgo. La validación cruzada y tests de permutaciones se utilizan para evaluar esta variable y el efecto combinado (Ritchie *et al.*, 2001).

El MDR se ha utilizado para detectar interacciones gen-gen en varios datos genéticos reales (Ritchie *et al.*, 2001; Moore y Williams, 2002; Julia *et al*., 2007), e incluso en presencia de errores de genotipado, datos incompletos, fenocopias y heterogeneidad genética (Ritchie *et al*., 2003b)

Aproximaciones Two Steps

Las aproximaciones *Two steps* (Hoh *et al.*, 2000) han sido ampliamente utilizadas como método *multi-marker* para el análisis de datos genéticos. Estas aproximaciones se denomina así porque constan de dos pasos:

i) selección de un número pequeño de marcadores potencialmente importantes.

ii) modelado de las interacciones entre los marcadores importantes y/o predictores medioambientales.

El segundo paso puede llevarse a cabo por los métodos estadísticos clásicos mencionados anteriormente.

Una aproximación *Two Steps* basada en técnicas de *bootstrapping* para la selección de marcadores fue utilizada por Hoh y colaboradores en un estudio con pacientes con una enfermedad de corazón (Hoh *et al.*, 2000). En un primer paso los autores calculan estadísticos para cada SNP, un Chi cuadrado a partir de una tabla de contingencia de 2x3 que corresponde a los 3 genotipos de casos y de controles. El efecto combinado de todos los marcadores se obtienen por la suma de todos los estadísticos. Para evaluar la significación estadística de esta suma se calcula el P valor a partir de un

determinado número de muestras *bootstrap* obtenidas bajo la hipótesis nula de no asociación. Ya que en la suma están contenidos todos los SNPs, incluidos muchos que no muestran asociación, este primer P valor es muy probable que no sea significativo, así que el SNP con estadístico menor se elimina de la suma. Este proceso se repite hasta obtenerse un P valor significativo, y los SNPs que permanecen en la suma se consideran preseleccionados.

En un segundo paso se crean réplicas al azar del conjunto de datos original, donde cada réplica es una muestra *bootstrap* obtenida bajo asociación. Para cada réplica se repite el procedimiento anterior, de forma que se obtiene para cada una de ellas un conjunto de marcadores preseleccionados.

Aquellos marcadores que han sido preseleccionados en más del 60% de las réplicas se consideran los marcadores importantes para la asociación a la enfermedad. Esta aproximación se aplicó con un grupo de 779 pacientes enfermos de corazón, 342 de los cuales desarrollaron restenosis (casos) y el resto no (controles) y se encontraron 11 marcadores, en 10 genes, asociados con la predisposición a sufrir restenosis (Hoh *et al.*, 2000). Este método aunque intuitivamente parece sencillo, no está puesto en un contexto de test de hipótesis estadístico ya que no se proporciona una significación estadística a la selección de SNPs.

Una modificación propuesta por los mismos autores, y conocida como *Set Assotiation Approach* (SAA) permite la selección de marcadores bajo control del nivel de significación (Hoh *et al.*, 2001).

En el SAA los estadísticos para cada SNP son el producto de dos Chi cuadrados, el primero de ellos lo calculan como se explicó anteriormente y mide la diferencia de frecuencias alélicas entre casos y controles. El segundo término mide la desviación del HWE en controles. El procedimiento es similar al anterior (figura 9), sólo que las sumas de los estadísticos se realizan por adición de los marcadores importantes, y no por eliminación de los no importantes.

Los marcadores se ordenan según su estadístico y se realizan las sumas incrementando el número de términos gradualmente desde 1 hasta un número máximo determinado (M). Por ejemplo S_3 será la suma de los 3 estadísticos mayores. Para cada S_i se calcula un nivel de significación estadística mediante un test de permutación.

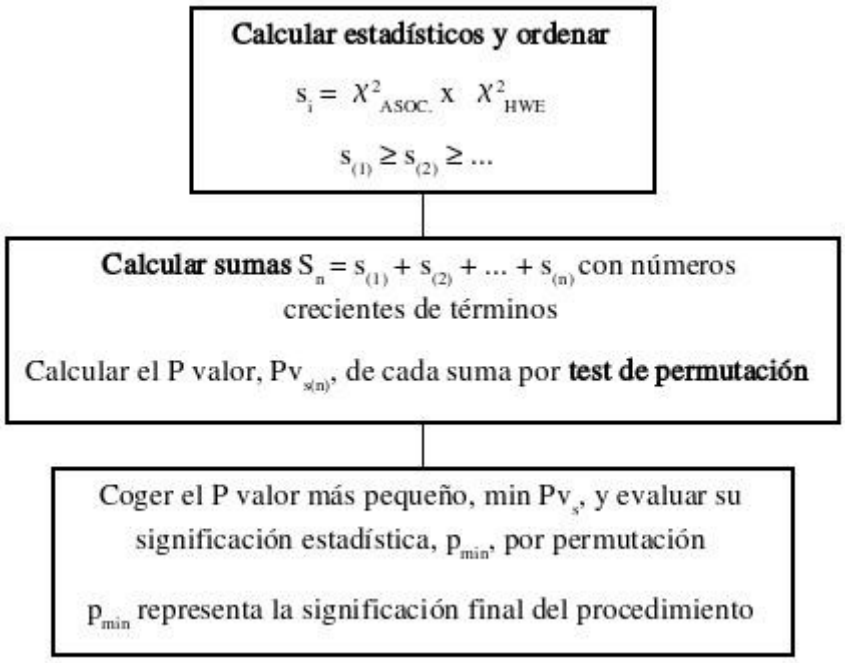


Figura 9. *Diagrama de flujo ilustrando el algoritmo implementado en la aproximación SAA.*

Las etiquetas de clases (casos y controles) se permutan al azar para conseguir un muestreo permutado donde no hay asociación. Se generan de esta forma muchos muestreos permutados de distintas sumas, y la proporción de muestras permutadas con una suma determinada que son mayores que la correspondiente suma origina es el P valor para esa suma.

De esta forma se genera un número M de P valores correspondientes a M sumas. El P valor más pequeño de todos se vuelve a evaluar por permutación para encontrar la significación estadística global. Los SNPs contenidos en la suma que proporcionó el P valor más pequeño son los seleccionados como asociados a la enfermedad.

Al aplicar el SAA al estudio anterior de Hoh (Hoh *et al.*, 2000), se seleccionaron 10 SNPs correspondientes a 9 genes, 6 de los cuales coinciden con los obtenidos bajo el procedimiento de *bootstrapping* y por tanto es probable que sean genes asociados a la enfermedad.

La aproximación SAA permite manejar un número elevado de marcadores, resuelve el problema de la dimensionalidad al reducirse ese número a un número menor de marcadores importantes, y además proporciona un nivel de significación general para los marcadores seleccionados. La mayor desventaja de este método consiste en que únicamente se testan las interacciones genéticas entre los marcadores

incluidos en la suma, con lo que pueden perderse interacciones importantes con efectos débiles que no lleguen a dar un estadístico elevado (Heidema *et al.*, 2006). Además este procedimiento no tiene en cuenta la correlación entre marcadores, algo que es importante sobre todo al manejar números elevados de SNPs.

Cuando muchos SNPs correlacionados del mismo gen se incluyen en la suma, el añadir SNPs a la suma basándose sólo en su estadístico puede ser ineficiente. Para poder manejar las correlaciones entre marcadores los autores propusieron otra modificación que ajusta los estadísticos de cada marcador considerando la correlación existente con los marcadores ya presentes en la suma (Wille *et al.*, 2003).

El método de SAA se ha implementado en dos programas, Sumstat y Statpval (http://www.genemapping.cn). El primero de ellos calcula la suma de estadísticos y el segundo evalúa el nivel de significación asociado al más pequeño de esos P valores.

El SAA puede manejar números grandes de marcadores y es útil para reducirlos a aquellos con una contribución importante en la enfermedad. Se centra por tanto en el primer paso de selección de marcadores, pero para el moldeado de interacciones gen-gen es necesario aplicarlo en combinación con otros métodos como MDR para detectar los genes importantes y las interacciones implicadas en las causas de la enfermedad.

2

OBJETIVOS

Como se ha comentado anteriormente, una forma de llevar a cabo estudios de asociación de una manera más efectiva es mediante la identificación de subgrupos de SNPs con alta probabilidad de conferir riesgo de desarrollar la enfermedad. En este sentido, es importante el desarrollo de estrategias *in silico* dirigidas a la predicción de SNPs con relevancia funcional.

Por otra parte, otro tipo de variación genómica, como son las variaciones en el número de copia, también pueden tener un efecto funcional, por lo que su análisis mediante *arrays* de CGH también debe ser tenido en cuenta.

Finalmente, la interpretación de los resultados obtenidos en los estudios de asociación no es a menudo sencilla. El elevado número de SNPs, sus complejas interacciones y sus distintas frecuencias poblacionales hacen necesario el desarrollo de métodos estadísticos y bioinformáticos que resuelvan estos problemas y además faciliten la interpretación de los resultados aportando un significado biológico.

La presente tesis ha querido contribuir a la resolución de los problemas antes citados mediante la consecución de los siguientes objetivos:

- El desarrollo de métodos de predicción del posible efecto funcional de SNPs a nivel transcripcional y su aplicación a la totalidad de polimorfismos identificados en el genoma humano.
- La creación y mantenimiento de herramientas bioinformáticas que permitan el acceso a esta información, y que integren, además, otros métodos de predicción de funcionalidad e información sobre frecuencias poblacionales y datos de LD para obtener un catálogo exhaustivo de marcadores genéticos con propiedades óptimas para genotipado
- El desarrollo y mantenimiento de una herramienta bioinformática para la visualización y detección de saltos en el número de copia para datos de aCGH
- El desarrollo de un método de análisis de datos de estudios de asociación en el que se integren diferentes fuentes de información biológica que facilite la interpretación de los resultados.

3

MATERIAL Y MÉTODOS

1. Selección de SNPs: PupaSuite

Con la idea de seleccionar conjuntos óptimos de SNPs usando toda la información sobre su posible efecto fenotípico, frecuencias poblacionales y LD, se han desarrollado varias aplicaciones web que finalmente se han integrado en una única herramienta llamada PupaSuite.

El núcleo de PupaSuite se ha desarrollado en Perl y javascript. Para aumentar el rendimiento de la herramienta todos los datos de SNPs y datos genómicos relacionados se precalculan y se guardan en diversas bases de datos gestionadas en MySQL.

1.1. Bases de datos y herramientas integradas en PupaSuite

La mayor parte de la información disponible en las bases de datos de PupaSuite, incluyendo secuencias genómicas, estructuras génicas, localización de SNPs, genes, tránscritos, alelos, validación, datos de frecuencia en distintas poblaciones, etc, se obtienen de la base de datos Ensembl instalada localmente.

Ensembl es un proyecto conjunto entre el European Bioinformatics Institute y el Sanger Institute. Aparece en 1999 y fue el primero en proporcionar un visor al borrador del genoma, curando a mano los resultados obtenidos a partir de análisis computacionales. Desde 2002 sus anotaciones se basan en los ensamblados del NCBI. Ensembl es una herramienta que permite la anotación y comparación de muchos genomas eucariotas, que además de dar información sobre los genes proporciona información sobre númerosas características genómicas como por ejemplo elementos repetitivos, citobandas, predicciones de islas CpG, regiones de homología con otras secuencias genómicas, etc. Además de sus propias anotaciones, Ensembl incorpora datos de otras bases de datos específicas como los genes asociados a enfermedad de OMIM, motivos de InterPro (http://www.ebi.ac.uk/interpro), anotaciones de Gene Ontolog (http://www.geneontology.org), predicciones de CisRed (http://www.cisred.org) y SNPs de la base de datos dbSNP. De dbSNP importa los datos de alelos, frecuencias y secuencias flanqueantes, y a partir de estos datos originales procesan otros como el cambio peptídico, localización genómica del SNP, etc.

Se decidió usar las anotaciones de Ensembl porque proporciona un acceso abierto a sus bases de datos en MysQL, tanto directamente como a través de su interfaz de programacion de aplicaciones

(API) asociada. Aunque Ensembl o dbSNP permiten la búsqueda de SNPs de acuerdo con un criterio específico (como localización genómica o frecuencia alélica) ninguna proporciona predicciones sobre las posibles consecuencias funcionales de los SNPs.

Bloques de haplotipos y LD

Los datos de genotipado se obtienen directamente de la base de datos del proyecto HapMap. Estos datos de genotipado se utilizan para el cálculo de los bloques y parámetros de LD. Para eso, una vez obtenidos los genotipos se corre la aplicación Haploview (Barrett *et al*., 2005) en java, que devuelve los bloques, haplotipos y tag SNPs para el conjunto de SNPs analizados.

Los haplotipos se estiman usando un algoritmo *Expectation Maximization* (EM) modificado para acelerar el algoritmo EM original, similar al método descrito por Qin y colaboradores (Qin *et al*., 2002). Los tagSNPs se seleccionan mediante una estrategia basada en el programa Tagger (de Bakker *et al*., 2005, 2006)

Regiones conservadas

Las regiones conservadas humano-ratón se obtienen directamente de Ensembl. El método utilizado por Ensembl es un análisis de los genomas completos mediante BLASTz (Schwartz *et al.*, 2003). Los datos obtenidos se procesan después para producir un subconjunto de regiones altamente conservadas usando el programa 'subsetAxt' (www.ensembl.org). Estas regiones conservadas y altamente conservadas pueden ser usadas para dar mayor verosimilitud a las predicciones de funcionalidad.

Otros programas y bases de datos utilizados en PupaSuite son Match™ (Kel *et al*., 2003), Transfac® (Wingender *et al*., 2000), Pmut (Ferrer-Costa *et al.*, 2002, 2004, 2005) y SNPeffect (Reumers *et al*., 2005, 2006).

- TRANSFAC® y Match™

  TRANSFAC® es una base de datos de elementos de DNA reguladores que actúan en cis y factores que actúan en trans para genes eucariotas, que contiene las matrices de pesos (PWMs) para los sitios de unión a esos factores. Match™ es una herramienta basada en PWMs que está interconectada y se distribuye con TRANSFAC®. Utiliza la librería de matrices de TRANSFAC® para identificar regiones promotoras en genes y localizar los elementos de secuencia consenso que puedan representar sitios de unión para factores de transcripción y

proporciona diferentes opciones para el filtrado de matrices o la disminución del número de falsos positivos.

- Pmut

Es un programa para la anotación y predicción de mutaciones patológicas que utiliza información basada en la secuencia (propiedades aminoacídicas e información evolutiva) y redes neurales para procesar esa información y determinar si un SNP codificante no sinónimo puede ser patológico o neutral.

- SNPeffect

La base de datos SNPeffect describe el efecto de SNPs codificantes no sinónimos en varias propiedades fenotípicas en proteínas humanas, usando herramientas bioinformáticas basadas en tanto en secuencia como en propiedades estructurales. Los fenotipos moleculares descritos se agrupan en tres categorías: estructura y dinámica, sitios funcionales y procesamiento celular. Entre las herramientas utilizadas en SNPeffect se encuentran herramientas desarrolladas en el propio grupo como FoldX (Fernandez-Escamilla *et al.*, 2004) que predice el cambio de estabilidad causado por el cambio de aminoácido, y Tango (Schymkowitz *et al.*, 2005), que predice regiones de agregación ß en la secuencia proteica, y otras herramientas externas como AmyScan (López de la Paz y Serrano, 2004), Psort II (Nakai y Horton, 1999), O-GlycBase (Gupta *et al.*, 1998) o Phosphobase (Kreegipuu *et al.*, 1999)

1.2. Búsqueda de SNPs con potencial efecto fenotípico

La aplicación web utiliza por tanto una base de datos precompilada, generada originalmente de datos genómicos procedentes de Ensembl, y que incluye información sobre el potencial efecto funcional de los SNPs, tanto a nivel transcripcional (alteración en el nivel de expresión o *splicing* alternativo) como a nivel de producto génico (alteraciones en la secuencia de la proteína). Para ello se busca información sobre SNPs que pudieran interrumpir sitios de unión a factores de transcripción, potenciadores de *splicing* exónicos, silenciadores de *splicing* exónicos, sitios canónicos de *splicing*, secuencias capaces de formar triple hélice y SNPs codificantes no-sinónimos con potencial efecto funcional.

1.2.1. SNPs en sitios de unión a factores de transcripción

Las secuencias de todos los genes contenidos en el genoma humano se obtienen de la base de datos Ensembl. Para cada gen se extrae la región 10,000pb río arriba del sitio de inicio de la transcripción (TSS) indicado por Ensembl, correspondiente a la región reguladora de los genes, buscando posibles sitios de unión a factores de transcripción. Aunque el escaneo se hace para la región 10,000pb río arriba de cada gen, el tamaño de la región a analizar puede ser modificada ya que es un parámetro de la herramienta.

Para la identificación de los sitios de unión a factores de transcripción (TFBSs) se utilizan las 358 matrices de pesos específicas de posiciones (PWMs) catalogadas en la base de datos Transfac® a través del programa Match™. Con el programa Match™ se utilizan solamente matrices de vertebrados de alta calidad y se selecciona un corte que minimiza el número de falsos positivos. Este corte se obtiene explorando las secuencias de los exones número 3 de cada gen para reducir el número de sitios obtenidos al azar por el programa (Kel *et al*., 2003).

Una vez identificados los TFBSs en las secuencias río arriba de cada gen, se buscan los SNPs situados en ellos y se anotan en la base de datos. Posteriormente se repite en análisis de las secuencias río arriba variando cada uno de los alelos de cada SNP situado en la secuencia, esté o no dentro de un TFBS. Los resultados se comparan con los obtenidos para las secuencias originales y de esta forma se anota para cada SNP si su presencia provoca la pérdida de un TFBS, si no le afecta o si genera uno nuevo.

1.2.2. SNPs en sitios de splicing

De Ensembl se obtiene la estructura exónica de todos los genes de humano y se localizan los dos nucleótidos conservados en las fronteras intrón-exón *y* que constituyen la señal de *splicing* (Cartegni *et al.*, 2002). Todos los SNPs situados en esas posiciones (y que por tanto podrían estar afectando al *splicing*) se guardan en la base de datos como SNPs con posible efecto funcional.

1.2.3. SNPs en potenciadores de splicing exónicos

Las secuencias exónicas de todos los genes humanos, incluyendo las zonas UTR, se escanean para predecir la presencia de potenciadores de *splicing* exónicos (ESEs) para las proteínas humanas SR (ricas en serina/arginina) siguientes: SF2/ASF, SC35, SRp40 y SRp55, por medio de las matrices de pesos disponibles para ellas (Cartegni *et al.*, 2003). Para cada sitio en el que se predice la presencia de un ESE (que sería un putativo sitio de unión para una proteína específica SR), se obtiene un *score* relacionado con la probabilidad de que ese sitio sea un ESE real. Sólo ESEs con *scores* mayores que un umbral mínimo son recogidos en el análisis (estos umbrales dependen de la proteína y son SF2/ASF: 1.956, SC35: 2.383, SRp40: 2.670 y SRp55: 2.676). Este umbral mínimo se establece como la mediana del *score* más alto para cada secuencia en un grupo de secuencias de 20 nucleótidos de longitud escogidas al azar entre el total de secuencias utilizadas inicialmente para la construcción de las matrices (Cartegni *et al.*, 2003).

Si un SNP cae en una de esas secuencias, el nuevo *score*, correspondiente a la secuencia mutada por el SNP, se vuelve a calcular. Si hay diferencias en los dos *scores* (por ejemplo, que con el SNP el *score* no llegue al umbral), se considera que ese SNP podría tener efecto en la regulación de los genes afectados, ya que podría estar inhibiéndo la acción de ese ESE. Las predicciones de SNPs en ESEs sólo se hicieron para SNPs bialélicos.

1.2.4. SNPs en silenciadores de splicing exónicos

Las secuencias exónicas obtenidas anteriormente para la búsqueda de ESEs se escanean de nuevo para la búsqueda de silenciadores (ESSs). Para esta búsqueda se utiliza el conjunto de ESSs candidatos (FAS-hex-3 set) obtenidos por Wang y colaboradores (Wang *et al.*, 2004). Todos los SNPs situados en esos motivos se guardan en la base de datos como SNPs que pudieran estar interrumpiendo la actividad silenciadora. Al no disponer de PWMs para los motivos ESS no se hacen predicciones sobre el efecto del alelo mutado, sino que solamente se señala la existencia de SNPs en los motivos. Para disminuir el número de falsos positivos la búsqueda puede hacerse teniendo en cuenta sólo los que aparecen en regiones conservadas humano-ratón.

1.2.5. SNPs en regiones capaces de formar triple hélice

Con el objetivo de detectar posibles SNPs que afecten regiones capaces de formar triple hélice (TTSs), se escanean las secuencias desde 10Kb río arriba hasta el extremo 3' de los genes, buscando secuencias de más de 10 polipurinas o polipirimidinas (putativos TTSs), y los SNPs localizados en esas regiones se guardan en la base de datos como SNPs con potencial efecto funcional.

1.2.6. SNPs codificantes no-sinónimos con putativo efecto patológico

Los SNPs que producen un cambio de amino ácido (nsSNPs) son probable que produzcan algún efecto fenotípico, y su putativo efecto patológico puede ser predicho por el algoritmo del programa Pmut (Ferrer-Costa *et al.*, 2002, 2004, 2005). Este algoritmo utiliza información basada en la secuencia (propiedades aminoacídicas e información evolutiva) y redes neuronales para procesar esa información y determinar cambios de aminoácido asociados a enfermedad. El servidor de Pmut implementa una pequeña red neuronal de 20 nodos y una capa oculta y tres descriptores derivados de secuencia (matrices de sustitución PAM40 y PSSM y un descriptor de variabilidad), que se obtienen de bases de datos o se derivan internamente de múltiples alineamientos utilizando PSI-Blast (Altschul *et al.*, 1997) sobre la base de datos no redundante de SwissProt/TrEMBL. Para obtener la predicción funcional de los nsSNPs se recogen todos los SNPs no sinónimos bialélicos de la base de datos Ensembl y se utiliza este algoritmo para clasificarlos como patológicos o neutrales y guardarlos en la base de datos.

El efecto de SNPs no-sinónimos también puede ser medido por medio de las propiedades físico-químicas y estructurales de las proteínas a las que afectan. La base de datos SNPeffect (Reumers *et al.*, 2005, 2006) utiliza distintas herramientas computacionales como Tango o FoldX para predecir cambios en el procesamiento celular, dinámica y estructura de la proteína. Los nsSNPs obtenidos a partir de Ensembl se cruzan con esta base de datos y las predicciones de SNPeffect se incorporan a la base de datos de Pupasuite.

Por último el efecto patológico de los polimorfismos puede ser estimado mediante estudios comparativos e información filogenética mediante el cálculo de la presión selectiva a nivel de codones (Arbiza *et al.*, 2006). Para cada uno de los SNPs no-sinónimos obtenidos de Ensembl se recoge su

posición, cambio de aminoácido y secuencia aminoacídica flanqueante y esos datos se utilizan para evaluar el posible efecto patológico de ese SNP mediante el método de Arbiza y colaboradores, que cuenta con dos aproximaciones alternativas: modelos de máxima verosimilitud basados en codones, implementados en PAML (Yang, 1997), y el método de *likelihood-ratio* (SLR) (Massingham y Goldman, 2005).

Ambas aproximaciones utilizan la comparación de tasas relativas de sustituciones sinónimas (dS) y no sinónimas (dN) para medir la presión selectiva como el cociente de estas tasas (ω = dN/dS). Si las mutaciones no sinónimas son dañinas, la selección positiva reducirá su tasa de fijación y ω será menor que 1, mientras que si las mutaciones no sinónimas son ventajosas, éstas se fijarán a una tasa más alta que las sinónimas y ω será mayor que 1. Una proporción de $\omega = 1$ es consistente con la evolución neutra.

De acuerdo con los autores del método, los codones con mutaciones que se encuentran frecuentemente asociadas a enfermedad tienen valores de ω menores de 0.1 y por tanto nsSNP localizados en esos codones tienen alta probabilidad de ser patológicos. La estima de presión selectiva se utiliza en PupaSuite para, según su valor de ω, predecir el potencial efecto patológico de todos los nsSNPs recogidos en Ensembl.

2. Análisis de variaciones de número de copia: ISACGH

Para el análisis de otro tipo de variación genómica, los CNVs, se ha creado otra herramienta web llamada ISACGH. Esta herramienta se ha desarrollado en el lenguaje de programación Perl y utiliza la base de datos de Ensembl y su API para localizar y representar gráficamente los datos de expresión génica o los datos de hibridación genómica introducidos por el usuario.

ISACGH recoge los datos procedentes de *microarrays* (clones, BACs u oligonucleótidos) y representa los valores de hibridación sobre sus correspondientes posiciones en el genoma, de acuerdo con las anotaciones de Ensembl. ISACGH acepta cualquier tipo de identificador para las sondas incluido en Ensembl.

Para la estimación de las regiones con variación en el número de copia la herramienta incorpora 4

métodos distintos. La base de estos algoritmos es indexar los datos (los datos son ratios, normalmente logaritmos, de las intensidades de hibridación que representan el número de copias) por la posición física de los clones en el genoma, para identificar regiones concentradas de ratios altos o bajos.

i) Método de *Smoothing*: es una variación del método *Adaptative Weights Smoothing* (Polzehl *et al*., 2000) y que ha sido implementado en el paquete GLAD (Hupe *et al*., 2004) de R. Esencialmente el algoritmo hace un suavizado de la linea de puntos y ajusta una función que cuantifica los saltos en la nube para identificar los bloques de diferente número de copia.

ii) Método de *Binary Segmentation*: este método evalúa si cada punto de los datos es un punto de corte. Esto lo hace de forma iterativa de forma que optimiza el orden en el cual se deben hacer las comparaciones (Olshen *et al*., 2004).

iii) Método de Regresión: este método es similar al método de S*moothing*. Ajusta una linea de regresión para cada N puntos consecutivos (por defecto N es igual a 10), para obtener un vector de pendientes equivalente de alguna forma a la curva derivada de los datos originales. Después se utiliza la estimación local de la variabilidad de los datos para identificar los picos de la pendiente que son suficientemente grandes como para indicar un punto de corte en los niveles de intensidad de hibridación.

iv) Método *Isowindow*: en este método se ordenan los puntos por su posición en el cromosoma, se selecciona un un determinado número N de puntos y se obtiene la media (u otro parámetro de centralización). Esto se repite para los N puntos situados a su derecha y a su izquierda y mediante un test de la t se obtiene el P valor que mide si hay una diferencia significativa entre los puntos que se están analizando y los de la vecindad. Si hay diferencia significativa los puntos analizados se toman como un bloque.

3. Análisis de datos de genotipado

Para el análisis de datos procedentes de estudios de asociación se ha desarrollado un programa que recoge los datos de genotipos de estudios de casos y controles y realiza distintos tests, algunos de los cuales incorporan información biológica obtenida de varias bases de datos, para obtener una selección de SNPs cuya significación estadística se mide mediante permutaciones y que se puede usar para

derivar una nueva hipótesis en el mecanismo de la enfermedad estudiada.

El programa está escrito en C e incorpora una base de datos precompilada con información sobre interacciones proteína-proteína, anotaciones de Gene Ontology y localización de SNPs en regiones conservadas obtenidas de Ensembl.

Interacciones proteína-proteína

Los datos sobre las interacciones proteína-proteína se obtienen de la base de datos BioGRID (Stark *et al.*, 2006). BioGRID es una base de datos pública de interacciones proteína-proteína para varias especies creada en 2003 como repositorio general de bases de datos de interacciones, y que por tanto incluye las anotaciones de otras bases de datos conocidas como BIND, DIP, MINT o MIPS (Stark *et al.*, 2006). BioGRID incluye anotaciones derivadas de técnicas de alto rendimiento como las técnicas de *Two Hybrid* o espectrometría de masas, y anotaciones derivadas de la literatura. En la versión 2.0.21 la base de datos contiene 38,223 anotaciones para humano.

Anotaciones de Gene Ontology (GO).

La base de datos de Gene Ontology (Ashburner *et al.*, 2000) contiene un vocabulario estructurado y controlado dividido en tres ontologías principales que describen productos génicos mediante términos referidos al proceso biológico, al componente celular y a la función molecular en los que el gen está implicado. El vocabulario de GO está estructurado en gráficos acíclicos dirigidos (DAGs) donde un vértice del gráfico corresponde a un término biológico y la unión con otro término muestra que esta relacionado con él. Los DAGs son similares a jerarquías pero difieren en que un término hijo (es decir, un término más especializado) puede tener muchos padres (términos menos especializados).

La base de datos de GO está incluida en la base de datos Ensembl, que es de donde se recogieron todos los datos de GO que utiliza el programa.

3.1. Método

En lineas generales, el método desarrollado es similar a las aproximaciones *Two Steps* mencionadas en la introducción, ya que en un primer paso se determina un subgrupo de marcadores que son importantes en la enfermedad de estudio y posteriormente, dependiendo del tipo de test elegido, se incorpora información funcional conocida *a priori* que sirve para dar más peso a aquellos marcadores con alguna evidencia de interacción biológica.

El programa incorpora 5 tipos de test: test de P valores, test de GO, test de interacciones proteína-proteína, test de conservación y test de Pritchard-Rosenberg. En todos estos test el primer paso es la preselección de un conjunto más pequeño de SNPs de entre todos los genotipados mediante un test de Chi cuadrado, después se obtiene un *score* asociado a ese grupo de marcadores, y posteriormente se testea la significación estadística del *score* mediante test de permutación (figura 10).

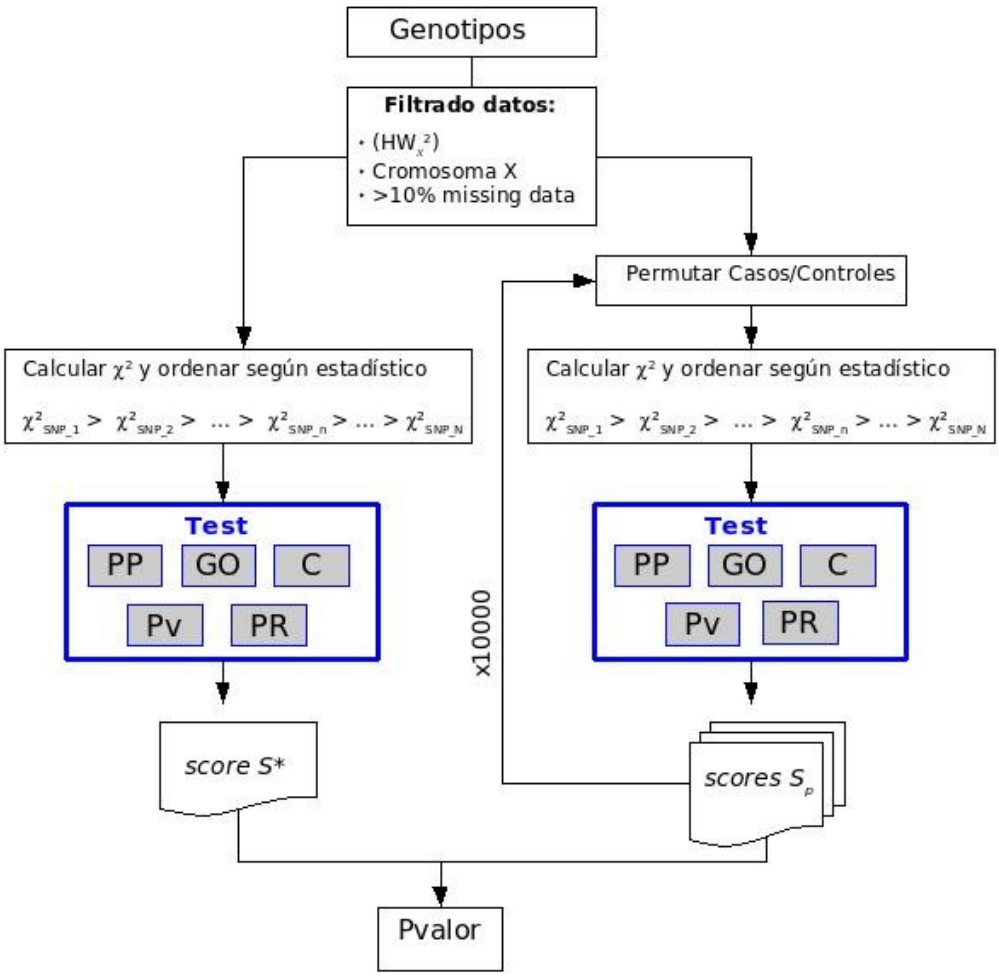


Figura 10. *Partiendo de los datos de genotipado (después de preprocesar para eliminar posibles errores de genotipado) de realiza un test de Chi cuadrado para ordenar los SNPs según su probabilidad de estar asociados a la enfermedad. Se selecciona un subconjunto de los mejores SNPs y se realiza un test (basado en interacciones proteína-proteína (PP), Gene Ontology (GO) o conservación (C)) que proporciona un S* funcional (izda). Esto se repite 10,000 veces permutando cada vez las etiquetas de casos y controles (dcha). El S* se compara con la distribución de 10,000 S_p para obtener una significación estadística. Además el programa incorpora otros tests para realizar análisis preliminares, como un test simple de P valores (Pv) y un test de Pritchard-Rosenberg (PR).*

Test de Chi cuadrado (χ^2).

Cuando lo que se pretende es comparar dos o más grupos de sujetos con respecto a una variable categórica, los resultados se suelen presentar a modo de tablas de doble entrada que reciben el nombre

de tablas de contingencia. En este caso, para testar la asociación alélica de cada SNP con la enfermedad se construye una tabla de contingencia 2x2 de frecuencias alélicas, donde cada celda es el número de veces que aparece un determinado alelo en los casos o en los controles (figura 11).

	Alelo1	Alelo2	Total
Casos	N_{11}	N_{12}	N_{1*}
Controles	N_{21}	N_{22}	N_{2*}
Total	N_{*1}	N_{*2}	N

Figura 11. *En la tabla, N_{11}, N_{12}, y N_{21}, N_{22} son las frecuencias observadas de los alelos 1 y 2 en casos y en controles respectivamente, siendo N el numero total de alelos en todos los individuos, y $N_{1*}=N_{11}+N_{12}$, $N_{2*}=N_{21}+N_{22}$, $N_{*1}=N_{11}+N_{21}$, y $N_{*2}=N_{12}+N_{22}$ los totales marginales.*

Para testar la diferencia alélica de cada SNP entre casos y controles, el valor de χ^2 se calcula:

$$\chi^2 = \sum_{j=1}^{r}\sum_{i=1}^{k}\frac{(N_{ij}-E_{ij})^2}{E_{ij}}, \quad donde \quad E_{ij}=\frac{N_{i*}N_{*j}}{N}$$

Bajo la hipótesis nula de independencia (el SNP no está asociado a la enfermedad), se sabe que los valores del estadístico χ^2 se distribuyen según la distribución de Chi cuadrado, que depende de los grados de libertad. Para el caso de una tabla de contingencia de r filas y k columnas, los grados de libertad son igual al producto (r-1)(k-1). Así, en el caso de la tabla anterior los grados de libertad son igual a 1.

De ser cierta la hipótesis nula, el valor obtenido debería estar dentro del rango de mayor probabilidad según la distribución de Chi cuadrado correspondiente. El P valor es la probabilidad de obtener los datos observados si fuese cierta la hipótesis de independencia. Así, para una seguridad del 99% ($\alpha = 0.01$) el valor teórico de una distribución Chi cuadrado con un grado de libertad es 6,63.

Test de permutación

A menudo, cuando el estadístico es complicado, no se conoce la distribución nula del estadístico como arriba. En vez de asumir una cierta distribución se puede construir una distribución nula

adecuada mediante la permutación de las etiquetas de clase, ya que de esta forma se garantiza la independencia (la no asociación). En nuestro caso se permutan las etiquetas de clase, es decir se permuta de forma aleatoria el estatus de enfermedad (casos y controles), y se recalculan los estadísticos.

Digamos que queremos testar si un *score* S^* que hemos calculado a partir de los datos observados es estadísticamente significativo. Entonces

i) se calcula S^* en los datos,

ii) para un número n de permutaciones, desde la permutación i=1 hasta n se permutan las etiquetas de fenotipo 'Y'

iii) para cada permutación i se calcula el *score* S_i

iv) se ordenan los *scores* S_i ... S_n y si nuestro S^* es más pequeño que el cuantil 0.05 o mayor que el cuantil 0.95, tendremos un resultados significativo a un nivel de significación del 5%.

Dependiendo del tipo de test y/o estadístico calculado podemos estar interesados solamente en S^* mayores o solamente en S^* menores que un cierto cuantil.

3.2. Test de P-valores (test PV)

En este test, en el primer paso se evalúa la asociación alélica de cada SNP individual. Para ello se construye una tabla de contingencia 2x2 de frecuencias alélicas. Para cada SNP se calcula el valor de χ^2 y el P-valor basado en la distribución χ^2 central bajo la hipótesis nula de no asociación con 1 grado de libertad. Para evitar la división por 0 cuando se calcula el estadístico se añade un valor positivo distinto de 0 a cada cuenta. Esto no debería tener un efecto significativo en la suposición de la distribución central de χ^2.

Se ordenan los SNPs según su estadístico de χ^2 y a partir de los estadísticos individuales se calcula un *score* global S^* para los primeros n SNPs de forma:

$$\text{Score}_{PV} = \sum_{i=1}^{n} \text{Pval}_i$$

Pval_i = Pvalor del SNP i

n = número de SNPs seleccionados

Después se permutan las etiquetas de clase de forma aleatoria. Bajo la hipótesis nula de no relación

alélica entre casos y controles, los *scores* S_i calculados a partir de esta distribución de permutaciones corresponderá a una distribución nula. Así se puede comparar el S^* original con la distribución nula para ver si existe significación estadística (figura 12). Al realizarse un único test se evita el problema del testeo múltiple.

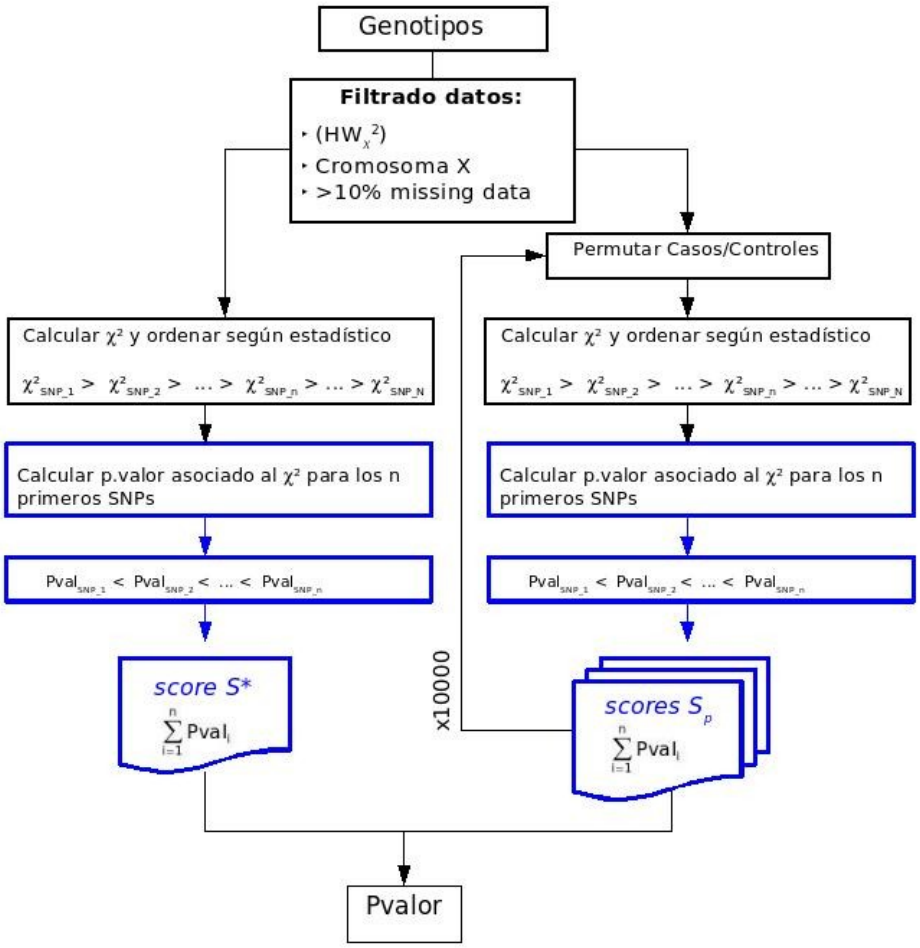


Figura 12. *Esquema test PV. En negro aparecen los pasos generales del método y en azul la parte especifica del test de P valores.*

3.3. Test de Gene Ontology (test GO)

Al igual que está generalmente aceptado que genes co-expresados juegan papeles comunes en la célula (Eisen *et al.*, 1998), pueden darse complejas interacciones entre marcadores que se pueden modelar teniendo en cuenta su localización en genes funcionalmente relacionados. En este test se incorpora información procedente de la base de datos de Gene Ontology (GO) con el objetivo de buscar grupos de SNPs relacionados funcionalmente y que como bloque presenten valores altos de

asociación. Con Gene Ontology se puede evaluar si un grupo de genes puede estar participando en algún proceso biológico común que podría entonces relacionarse con la enfermedad estudiada.

Gene Ontology proporciona anotaciones para genes, por lo tanto para asociar anotaciones a SNPs se necesita hacer primero una asociación de SNPs a genes. Para hacer esta asociación SNP-GO se utiliza la base de datos de Ensembl para recoger genes asociados a esos SNPs y los datos de LD de HapMap. Se utilizan los genotipos de población africana para ser más conservadores, ya que el grado de LD en población africana es más pequeño que en otras poblaciones.

A cada SNP se le asocia un gen si *i)* el SNP está dentro del gen o *ii)* el SNP está en LD ($r^2 > 0.9$) y a menos de 20Kb de otro SNP localizado en el gen.

Los SNPs se ordenan según su estadístico de χ^2 y para el conjunto de los mejores 'm' SNPs se obtiene el conjunto 'n' de genes asociados g_i, i={1,...,n}. A cada gen se le asocia el valor del χ^2 obtenido para el SNP al que está asociado. Si un gen aparece asociado a más de un SNP, el χ^2 que se le asigna es el mayor de todos. Así se evita la repetición de un gen asociado a muchos SNPs debido al LD entre ellos, condicionado por los marcadores seleccionados previamente.

Se hacen todas las posibles combinaciones de pares de genes i,j={1,...,n} i≠j, y para cada par se calcula un estadístico que es el producto de dos términos, el primero es la suma de los χ^2 de los genes del par, y el segundo un término que mide el nivel más especifico al que los genes del par comparten un término GO. El *score* global *S** se calcula como el sumatorio para todos los pares:

$$\text{Score}_{GO} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(\chi_i^2+\chi_j^2)\text{GO}_{niv}}{n}$$

n = número de genes
χ_i^2 = Chi cuadrado del Gen i
χ_j^2 = Chi cuadrado del Gen j
GO_{niv} = Nivel de GO más alto compartido entre el par de genes i y j

Se divide por 'n' para normalizar los *scores* ya que se pueden obtener diferente número de genes a partir de distintos conjuntos de SNPs de un mismo tamaño.

Si dos genes (dos SNPs) están altamente asociados con la enfermedad, entonces su estadístico será alto. Si esos genes están asociados en algún proceso o función específicos, entonces elevarán esa interacción particular. Si en cambio comparten solamente funciones más generales, el peso que se le da a esa interacción será menor. Finalmente si no comparten ninguna función, el estadístico será igual

a cero. Nuestra suposición es que no esperamos que SNPs al azar, es decir, no asociados, compartan GOs muy específicos y por tanto le daremos pesos pequeños.

Una vez obtenido el *score S**, se permutan las etiquetas de casos y controles de forma aleatoria, se recalculan los χ^2 y se obtiene un nuevo conjunto 'n' de genes asociados a los mejores 'm' SNPs. Como antes, se buscan los GOs comunes y se calculan nuevos *scores* S_i para comparar *S** con esta distribución nula (figura 13). Mediante el test de permutación se mantiene la dependencia entre los SNPs y no es necesario asumir ningún tipo de distribución *a priori*.

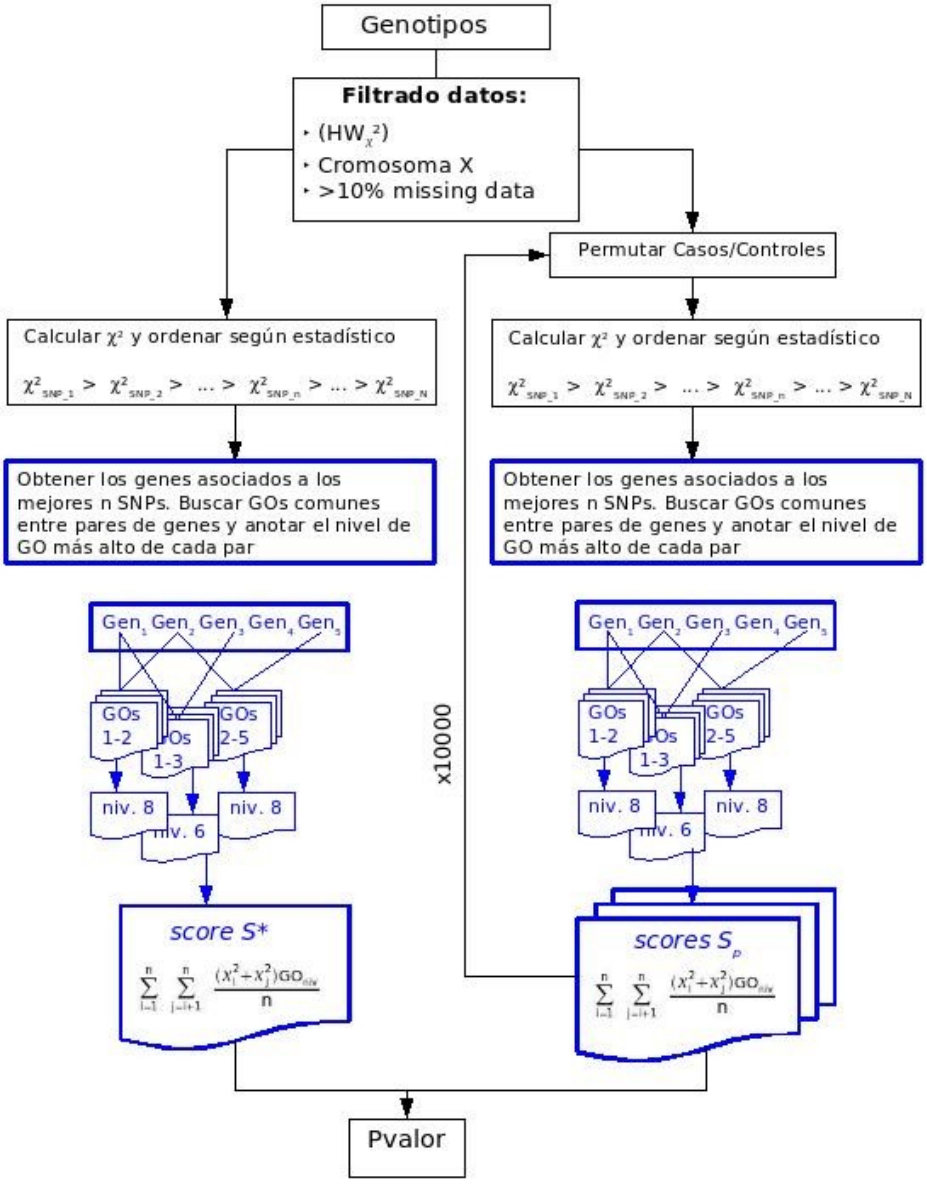


Figura 13. *Esquema test GO. En negro aparecen los pasos generales del método y en azul la parte especifica del test de GO*

El concepto básico de este test es calcular un *score* que combina información para muchos SNPs

sumando los estadísticos individuales que miden su asociación a la enfermedad, pero dando más peso a aquellos pares que interaccionan, de forma indirecta, a través de una asociación funcional.

3.4. Test de interacción proteína-proteína(test PP)

Este test se centra en SNPs que pertenecen a genes o proteínas que se conoce que interactúan en alguna red biológica. Al igual que en el test de GO, sólo se consideran los pares de genes que interaccionan, pero esta vez se consideran interacciones directas a través de una asociación física.

El procedimiento es el mismo que el del test GO, pero para cada par de genes el estadístico que se calcula es el producto de la suma de los χ^2 de los genes del par por un factor PP igual a 1 si hay descrita una interacción proteína-proteína para ese par, o igual a 0 si no hay descrita esa interacción.

$$Score_{PP} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(X_i^2 + X_j^2) \cdot PP}{n}$$

n = número de genes

X_i^2 = Chi cuadrado del Gen i

X_j^2 = Chi cuadrado del Gen j

PP = 1 si el par de genes interacciona; PP = 0 si no

La significación estadística del *score* se obtiene mediante el test de permutación de igual forma que en el test de GO.

3.5. Test de conservación (test C)

En este test no se tienen en cuenta interacciones sino que el *score* se calcula como el sumatorio de los estadísticos individuales de cada SNP, pero sólo para aquellos situados en zonas conservadas del genoma. La idea se basa en la observación de que el grado de conservación puede indicar cuales son las posiciones que tienen mayor probabilidad de estar asociadas a enfermedad (Mooney *et al.*, 2003, Mooney, 2005).

En primer lugar se calcula el estadístico individual χ^2 para cada SNP y este estadístico se utiliza para ordenar los SNPs por su asociación a la enfermedad. Para los mejores n SNPs se obtiene el *score* *S** como el sumatorio de sus estadísticos multiplicados por un factor de conservación, que es igual a 1

si el SNP esta situado en una zona conservada, o igual a 0 en caso contrario.

$$Score_C = \sum_{i=1}^{n} \chi_i^2 \cdot C$$

n = número de genes

χ_i^2 = Chi cuadrado del SNP i

C = 1 si el SNP está en una zona conservada; C = 0 si no

La significación estadística del *score* se obtiene mediante el test de permutación de igual forma que en los test anteriores.

3.6. Test de Pritchard-Rosenberg (test PR)

El test de Pritchard y Rosenberg (Pritchard y Rosenberg, 1999) evalúa las posibles asociaciones falsas que resultan de la presencia de estratificación de poblaciones en los datos. Las asociaciones falsas (es decir, las que no tienen un significado biológico, aunque puedan tener un significado genético real debido a la mezcla de poblaciones) pueden ocurrir si i) las frecuencias alélicas de los marcadores de interés difieren en las poblaciones que forman las muestras de estudio, y ii) si el riesgo a desarrollar el fenotipo también difiere por población. Pritchard y Rosenberg (Pritchard y Rosenberg, 1999) describieron un método para testar la estratificación mediante el uso de un estadístico que se define como la suma de los estadísticos χ^2 obtenidos al comparar frecuencias alélicas entre casos y controles para un grupo de marcadores no ligados, con grados de libertad igual a la suma de grados de libertad de todos los marcadores.

Este test se implementó en el programa para poder testar la posible estratificación poblacional. El poder para detectarla dependerá del número de marcadores utilizados en el test, por lo que es importante elegir un número lo suficientemente grande para asegurarse de que el test tiene suficiente poder para detectar una estratificación moderada.

En este test se seleccionan un número 'n' de SNPs al azar y se construyen sus tablas de contingencia con sus datos de frecuencias alélica en casos y controles. Se calculan los 'n' estadísticos χ^2 individuales y posteriormente se calcula un estadístico global χ^2_S como la suma de los estadísticos individuales.

$$\chi^2_S = \sum_{i=1}^{n} \chi^2_i$$

Bajo la hipótesis nula de no asociación, el estadístico global χ^2_S sigue una distribución de Chi cuadrado con n grados de libertad. Al estar los marcadores elegidos al azar, es improbable que estén ligados al locus de enfermedad, y por tanto una asociación significativa mostraría que existe estratificación.

4

RESULTADOS

1. SNPs con posible efecto funcional

1.1. SNPs situados en TFBSs

Se ha mostrado que el *score* de un TFBS para un TF determinado, obtenido mediante matrices de pesos de posiciones (PWM), construidas a partir de una colección de sus sitios de unión conocidos, puede proporcionar una estimación bastante ajustada de la afinidad de unión de ese TF a ese sitio *in vitro* (Stormo, 2000). Esta observación y los principios de termodinámica que hay detrás forman la base de la mayor parte de programas bioinformáticos genéricos usados para predecir TFBSs en DNA genómico (GuhaThakurta, 2006). Los eventos de unión TF-DNA *in vivo* son mucho más complejos ya que esta unión depende del contexto (por ejemplo de otras uniones cercanas, de la estructura local del DNA, etc.). Desafortunadamente esta información contextual sólo está disponible en raras ocasiones, de modo que no pueden usarse generalmente para la predicción de TFBSs. Por tanto, aunque el cambio en el *score* del TFBS puede no ser un predictor preciso de la unión del TF a su sitio de unión al DNA *in vivo*, en ausencia de otra información específica, la aproximación que hemos tomado aquí es una estrategia razonable para examinar el posible efecto del SNP sobre la unión del TF al TFBS.

En el presente trabajo se analizaron un total de 31,714 genes humanos anotados en Ensembl (versión 39.36a), correspondiente al ensamblado 36 del NCBI, y que contiene la versión 125 de dbSNP. Para cada gen se obtuvo la secuencia en formato FASTA de la zona promotora 5Kb río arriba del TSS indicado por Ensembl.

La decisión sobre en que región buscar TFBSs es, en cierta manera, arbitraria. La aproximación tomada aquí fue determinar TFBSs en los 5Kb precedentes al gen, ya que, aunque los elementos reguladores de la transcripción están a menudo enriquecidos en la región promotora inmediata (Montgomery *et al.*, 2007), también se pueden extender distancias más largas, a veces más de 100 Kb (Loots *et al.*, 2000).

En este estudio las predicciones de TFBSs se hicieron con las PWMs que representan los sitios de unión al DNA de los TF disponibles en la base de datos de TRANSFAC® mediante el programa Match™. Sólo se usaron PWMs generadas partir de la colección de sitios de unión al DNA de vertebrados. En el momento del análisis sólo había modelos de sitios de unión para 358 TFs de vertebrados, mientras que el número de distintos TFs en mamíferos se ha estimado en

aproximadamente 2,000 (Waterston *et al,* 2002), por tanto la predicción de sitios de unión al DNA de la mayoría de TFs no fue posible.

Después de mapear los SNPs en las regiones promotoras se encontraron 147,825 posibles TFBSs interrumpidos por un total de 95,255 SNPs (casi un 1% del total de SNPs en el genoma humano). De los casi 32,000 genes, un total de 28,067 presentaron al menos una predicción de TFBS interrumpido por un SNP, lo que constituye una proporción considerable (un 88%) del número total de genes. Sin embargo que un SNP esté en un TFBS no implica que tenga un efecto sobre su función. Para saber cuántos SNPs pueden estar afectando realmente al TFBS, se obtuvieron todos los SNPs de las secuencias promotoras.

Para cada SNP se tomó como alelo normal el coincidente con la secuencia original (ensamblado 36 del NCBI), y como mutantes el resto de alelos (uno normalmente). En cada secuencia promotora, para cada alelo mutante de cada SNP se generó una nueva secuencia FASTA con solamente ese cambio de alelo sobre la cual se volvió a ejecutar el programa Match™. De esta forma se cuantifican las diferencias en el *score* predicho para un TFBS con el SNP respecto al alelo normal. También así se pueden detectar pérdidas completas del TFBS o incluso la aparición de nuevos TFBSs debido a la variación polimórfica.

De entre los SNPs localizados en TFBSs, 24,241 SNPs (un 25.44%) obtuvieron un **cambio en el *score*** para el alelo mutante. Las diferencias en los *scores* variaron en el rango 0.001-0.12, siendo mayor de 0.5 en 1,380 ocasiones.

Se encontraron 43,850 SNPs que podrían haber generado **nuevos TFBSs,** ya que con los alelos mutantes el programa Match™ predice TFBSs que no ocurren en la secuencia original. Entre ellos, 6,461 SNPs generan un *score* perfecto (*matrix score* = 1) con el nuevo TFBS.

Además, 38,547 SNPs (40.46% de los SNPs en TFBSs) podrían estar provocando la **pérdida del TFBS**, ya que motivos que se reconocen con el alelo normal, no son reconocidos por el programa Match™ cuando se cambia al alelo mutante. Incluso en 6,321 casos, los motivos que se pierden se predecían con *scores* perfectos (*matrix score* = 1) en la secuencia original.

Por último se encontraron 6,429 SNPs (6.74%) cuyo cambio de alelo no produce **ningún cambio** de *score* para los TFBSs que se detectaron en la secuencia original. Estos suelen ser SNPs que caen en TFBSs muy degenerados, como los sitios de unión al factor de transcripción HNF-1, para el que aparecen descritas en TRANSFAC® tres matrices distintas que identifican miles de putativos sitios de unión para ese factor.

Ya que los TFBSs son típicamente cortos y degenerados, las predicciones que se obtienen con PWMs suelen contener un alto porcentaje de falsos positivos. Para incrementar la especificidad de los elementos TFBSs encontrados se adoptó un criterio de conservación.

Para ello se buscaron todas las regiones conservadas humano-ratón en las secuencias promotoras de todos los genes, y se anotaron aquellos TFBSs predichos que caen en regiones conservadas. Si se tiene en cuenta solamente estos, el número de SNPs predichos con posible efecto funcional se reduce considerablemente (tabla 1).

	Secuencias promotoras	*Regiones conservadas de secuencias promotoras*
# TFBSs predichos	1,968,274	586,172
# SNPs en TFBSs (% del total de SNPs)	95,255 (0.91%)	31,823 (0.30%)
# Genes afectados (% del total de genes)	28,067 (88%)	15,898 (50%)
SNPs en TFBSs	***Secuencias promotoras (% de SNPs en TFBSs)***	***Regiones conservadas de secuencias promotoras (% de SNPs en TFBSs)***
# SNPs que producen cambio de *score* (> 0.05)	1,380 (1.44%)	440 (1.38%)
# SNPs que generan nuevos TFBSs	43,850	13,978
# SNPs que generan nuevos TFBSs con *matrix score* = 1	6,461	2,156
# SNPs que provocan pérdida de TFBS	38,547 (40.46%)	12,910 (40.56%)
# SNPs que provocan pérdida de TFBS con *matrix score* = 1	6,321 (6.63%)	2,236 (7.02%)
# SNPs que no producen cambio	6,429 (6.74%)	2,002 (6.29%)

Tabla 1*. SNPs situados en TFBSs. El 88% de los genes humanos podrían estar afectados por la presencia de un SNP en TFBSs cercanos. Este porcentaje se reduce al 50% si solo consideramos TFBSs localizados en regiones conservadas.*

Debido a la existencia de numerosos TFBSs con secuencias consenso muy similares, a menudo se obtienen predicciones múltiples para una misma localización genómica. Como resultado se encuentran casos de SNPs que producen pérdidas, ganancias y cambios de *score* para distintos TFBSs a la vez. Por ejemplo, el SNP rs7068288 (A/G), situado en la posición -190 con respecto al gen YME1L1. Como se observa en la tabla 2, el SNP se encuentra en una zona de alta repetición A/T, que es una zona consenso para la unión de varios TFBSs.

TFBS	*Secuencia*	*Posición TFBS*	*Score Matriz*	*Efecto*
CDX	aat**a**cataaataaATAAA	-193 -176	0.924	nuevo *score*: 0.916
POU1F1	aT**G**CATaaat	-192 -183	0.991	nuevo TFBS
HNF1	cactaagaaaa**A**TTAAaaata	-201 -184	0.861	sin cambio
CDX	taaaaat**a**cataaATAAA	-197 -180	0.886	pérdida TFBS

Tabla 2*. El SNP rs7068288 se encuentra en una zona de repetición A/T, en la que se predicen cuatro TF de secuencia consenso parecida. El cambio de alelo A->G podría producir la pérdida de la unión del factor de transcripción CDX y podría hacer que el factor de transcripción POU1F1 se uniera a la zona promotora del gen YME1L1.*

1.2. SNPs situados en sitios de *splicing*

Como se ha visto el paso de la eliminación de los intrones es crítico en el proceso de *splicing*, un proceso que requiere un reconocimiento de los sitios donador y aceptor por la maquinaria de *splicing*. Aunque estos sitios de *splicing* (dinucleótidos GT y AG) no sean por si solos suficientes para el proceso, son necesarios, y la mayoría de los polimorfismos de una base situados en sitios de *splicing* que causan enfermedades ocurren en esas posiciones (Baralle y Baralle, 2005)

Para estudiar los posibles efectos de los SNPs sobre el proceso de *splicing* de los 31,714 genes humanos se obtuvieron las estructuras génicas y se localizaron los sitios donador y aceptor de cada unos de los intrones. Se encontraron un total de 1,122 SNPs (0.01% del total de SNPs) situados en estas zonas, afectando a un total de 2,043 tránscritos de 1,058 genes distintos (3.33% del total).

1.3. SNPs situados en ESEs

Para analizar la relación entre las variaciones genéticas simples en humanos y ESEs, se obtuvieron las secuencias (incluyendo las secuencias UTR) de los 282,420 exones anotados en Ensembl v39.

Utilizando las matrices de pesos disponibles para las 4 proteínas SR humanas SF2/ASF, SC35, SRp40 y SRp55, se escanearon las secuencias y se encontraron más de 12 millones de motivos ESE con *scores* significativos (*scores* mínimos en el rango de 1.956 a 2.676 dependiendo de la proteína, Cartegni *et al.*, 2003). Sin embargo, a pesar de la cantidad de motivos encontrados, sólo en un 4% de estos motivos aparece solapado un SNP.

Se encontraron un total de 223,487 SNPs en exones. Entre estos SNPs codificantes, hay 133,644 (1.28% del total de SNPs) SNPs diferentes localizados en los putativos motivos ESE encontrados, aunque solamente 91,613 de estos SNPs (0.87% del total) presentan una diferencia de *scores* (para los alelos mutante y normal) suficiente como para predecir una pérdida de la actividad del ESE. Estos SNPs potencialmente funcionales afectarían a un total de 20,997 genes distintos (66% del total de genes).

Si tenemos en cuenta sólo los ESEs localizados en regiones conservadas humano-ratón, las predicciones se reducen, aunque en menor grado que el observado con TFBSs, ya que precisamente las regiones conservadas suelen estar localizadas en las zonas codificantes (tabla 3)

	Secuencias exónicas	*Regiones conservadas de secuencias exónicas*
# ESEs predichos	> $12x10^6$	> $3x10^6$
# SNPs en ESEs (% del total de SNPs)	133,644 (1.28%)	112,104 (1.07%)
# SNPs que producen cambio de *score* (% del total de SNPs)	91,613 (0.88%)	76,828 (0.73%)

***Tabla 3**. Resultados de SNPs situados en ESEs. Un 0.88% de los SNPs de humano están situados en los ESEs predichos con las matrices disponibles para las 4 proteínas SR humanas SF2/ASF, SC35, SRp40 y SRp55. El porcentaje es ligeramente menor, 0.73%, cuando sólo se consideran predicciones de ESEs en regiones conservadas.*

Analizando la posición de los ESEs dentro de los exones, se puede ver que la densidad de ESEs no es uniforme a lo largo del exón, con la mayor densidad de ESE en las zonas cercanas a los sitios de *splicing* 5' y 3' y una menor intensidad a medida que se alejan hacia la zona interna (figura 14).

Sin embargo la distribución de los SNPs exónicos no parece ser así. Aunque no se aprecia una tendencia opuesta tan clara como se ha descrito previamente (Fairbrother *et al.*, 2004), sí se observa cierto aumento en el número de SNPs a medida que aumenta la distancia a los sitios de *splicing* (figura 15). Debido la mayor densidad de ESEs cerca de los sitios de *splicing*, las mutaciones que interrumpan la actividad del ESE son más susceptibles de ser eliminadas por selección purificadora. Esto podría potencialmente explicar la tendencia de la densidad de SNPs opuesta descrita anteriormente (Fairbrother *et al.*, 2004).

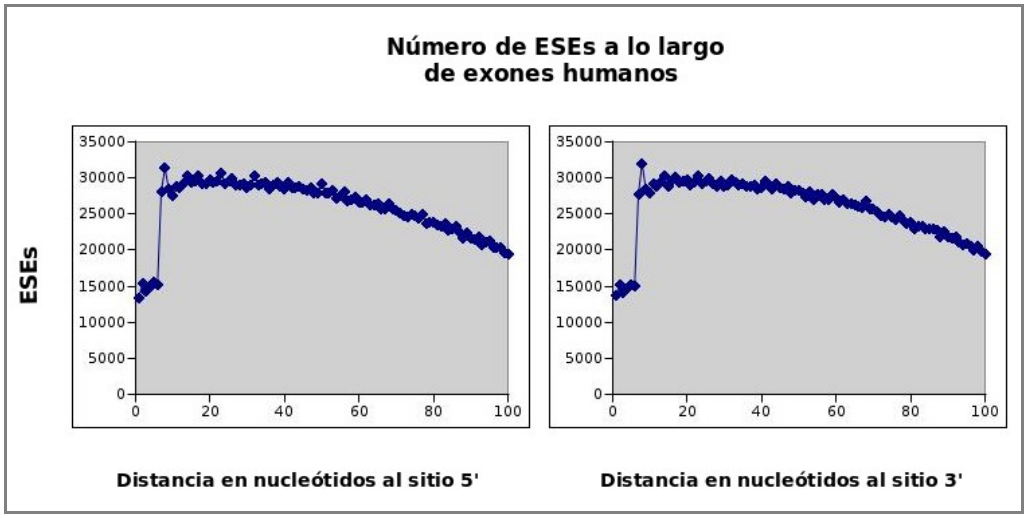


***Figura 14**. Número de ESEs a los largo de exones. El eje de abscisas muestra la posición de inicio de cada ESE con respecto a los sitios de splicing 5' (**izda**) y 3' (**dcha**). La gráfica muestra que los ESEs aparecen con mayor frecuencia en las zonas cercanas a los sitios de splicing y en menor frecuencia a mediada que se alejan hacia la zona interna del exón.*

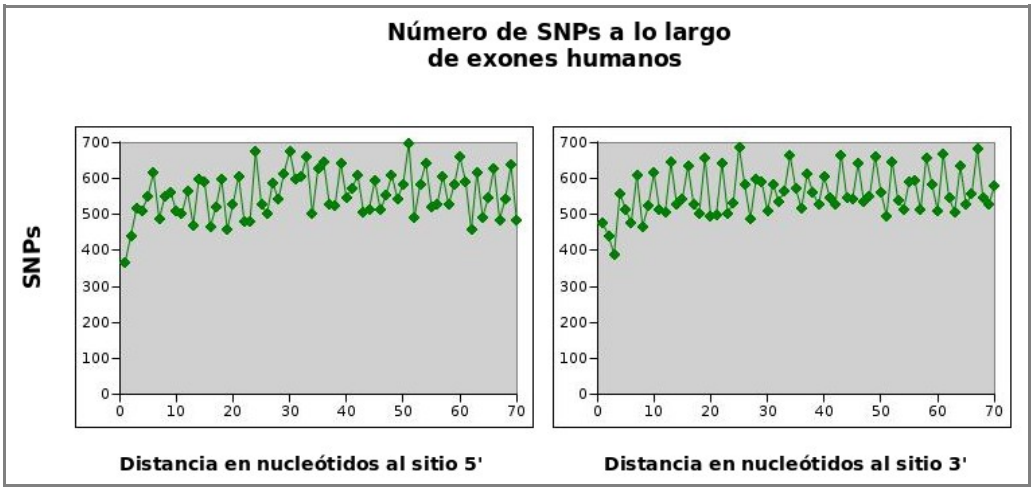


***Figura 15**. Número de SNPs a los largo de exones. El eje de abscisas muestra la posición de inicio de cada SNP con respecto a los sitios de splicing 5' (**izda**) y 3' (**dcha**). La gráfica muestra que el número de SNPs tiene una ligera tendencia a aumentar a medida que se alejan de los sitios de splicing.*

1.4. SNPs situados en ESSs

Aunque la cifra de 12 millones de ESEs encontrados supone que se han encontrado putativos ESEs para todos los genes, no quiere decir que todos esos potenciadores estén funcionando realmente como ESEs, ya que silenciadores próximos pueden hacer que la proteína SR no se una al motivo. Para comprobar la distribución de ESSs en el genoma y ver su relación con los ESEs y SNPs, se tomaron

las secuencias exónicas anteriores y se escanearon buscando los 103 motivos ESS candidatos de los que se dispone su secuencia (Wang *et al.*, 2004).

Se han encontrado un total de 1,852,396 ESSs, aunque sólo hay 17,957 SNPs en estos motivos, por tanto, al igual que con los ESEs, el porcentaje de motivos ESS que solapan con SNPs es muy pequeño. Si se analiza la distribución de silenciadores, observamos que estos elementos tienen una distribución en exones similar a la de ESEs (figura 16), concentrándose mayoritariamente en las zonas cercanas a los sitios de *splicing*. Esta coincidencia de ESEs y ESSs en la misma región exónica puede hacer que unos anulen la función de los otros, o como se ha descrito, esas zonas pueden ser nuevos elementos reguladores, donde existen funciones potenciadoras y silenciadoras solapadas que no son completamente dependientes de la unión de proteínas SR (Pagani *et al.*, 2003). El efecto funcional de los SNPs situados en estas zonas no se puede predecir por medio de las matrices de proteínas SR, y aunque podrían ser necesarios ensayos funcionales que corroboraran su efecto en el proceso de *splicing*, estos SNPs no deberían ser ignorados a la hora de preseleccionar SNPs con potencial efecto funcional.

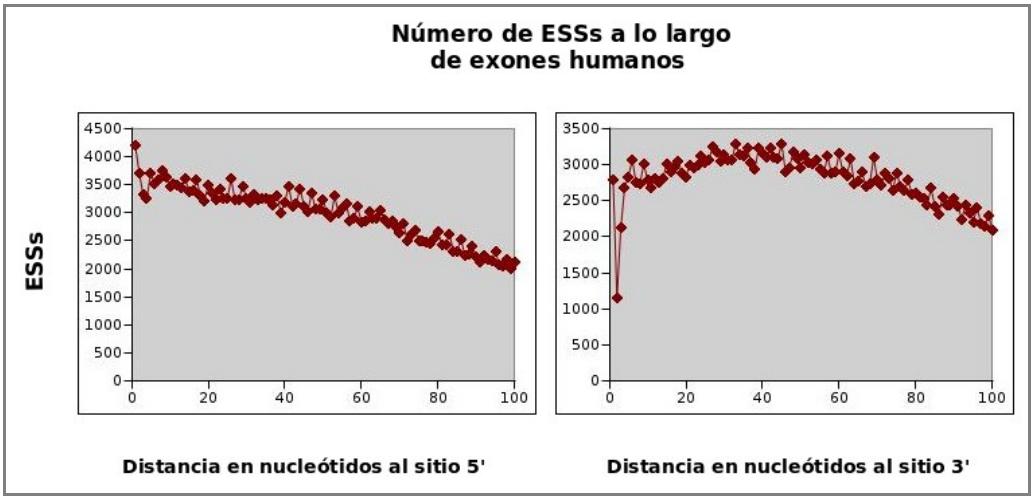


***Figura 16**. Número de ESSs a los largo de exones. El eje de abscisas muestra la posición de inicio de cada ESS con respecto a los sitios de splicing 5' (**izda**) y 3' (**dcha**). La gráfica muestra que los ESEs tienden a concentrarse en las zonas cercanas a los sitios de splicing y su densidad disminuye a medida que se alejan de los extremos del exón.*

1.5. SNPs situados en TTSs

Los tríplex de DNA (Pauling y Corey, 1953; Felsenfeld *et al.*, 1957) se han propuesto como regiones reguladoras para el control de la expresión génica (Goñi *et al.*, 2004). Las secuencias capaces

de formar triple hélice (TTSs) son secuencias de más de 10 polipurinas o polipirimidinas, y los SNPs localizados en esas secuencias podrían afectar la formación del tríplex y por tanto interrumpir la regulación normal de un determinado gen.

Para detectar estos posibles SNPs funcionales, se escanearon las secuencias de todos los genes del genoma, desde la posición 5Kb río arriba hasta el extremo 3' de cada gen. En estas secuencias se buscaron todos los putativos TTSs (polipurinas o polipirimidinas con una longitud mínima de 10 nucleótidos), y se mapearon todos los SNPs localizados en ellas.

Se encontraron más de 5 millones de TTSs en las secuencias analizadas. En estas secuencias se localizaron un total de 299,947 SNPs (2.87% del total de SNPs), estando la mayoría (270,569 SNPs) en la parte génica de las secuencias.

Esta cifra también se reduce considerablemente (47,549 SNPs, un 0.45% del total) al considerar únicamente aquellos que se sitúan en regiones conservadas.

1.6. Casi 500,000 SNPs con posible efecto regulador

Los métodos anteriores se han utilizado para identificar y evaluar sistemáticamente SNPs con potencial efecto regulador en el genoma humano. De los 10,430,753 SNPs contenidos en la versión 125 de dbSNP (incluida en la versión 39 de Ensembl), se han encontrado un total de 499,640 SNPs en regiones con importancia en regulación como TFBSs, ESEs, ESSs, TTSs y sitios de *splicing*. Esto supone que casi un 5% de los SNPs del genoma humano podrían tener una posible relevancia funcional en el genoma humano.

Una característica que puede ser indicativa de las consecuencias dañinas de un alelo es la evidencia de selección purificadora (Zhao *et al.*, 2003). La selección purificadora es la forma natural de selección que actúa para eliminar selectivamente mutaciones dañinas. Para comprobar si hay evidencia de selección purificadora se buscaron, para todos los SNPs anotados como funcionales en las categorías anteriores, los datos de frecuencia del alelo minoritario (MAF) para las cuatro poblaciones de HapMap. Estos datos se buscaron también para el resto de SNPs situados en las mismas zonas que no se predijeron como funcionales (tabla 4).

Tipo SNP	*Código*	*Número SNPS*	*media MAF CEU*	*media MAF CHB*	*media MAF JPT*	*media MAF YRI*
Sitios de *splicing*	SP	1,122	0.047268	0.048400	0.050595	0.047870
ESSs	ESS	17,957	0.124450	0.115093	0.115757	0.128984
pérdida ESEs	pESE	91,613	0.118910	0.112145	0.112366	0.124283
pérdida ESEs conserv.	pESEc	76,828	0.116024	0.109717	0.109830	0.121176
TTSs	TTS	299,947	0.141657	0.131273	0.130902	0.150613
TTSs conserv.	TTSc	47,549	0.139841	0.128183	0.127847	0.145238
nuevos TFBS	nTFBS	43,850	0.143668	0.132240	0.131360	0.152110
nuevos TFBS conserv.	nTFBSc	13,978	0.141991	0.130809	0.129825	0.148607
pérdida TFBS	pTFBS	38,547	0.147128	0.133775	0.133442	0.156299
pérdida TFBS conserv.	pTFBSc	12,910	0.148071	0.131882	0.131761	0.153759
cambio *score* TFBS (dif. *scores* > 0.05)	cTFBS	1,380	0.141089	0.122893	0.118766	0.148988
cambio *score* TFBS conserv. (dif. *scores* >0.05)	cTFBSc	440	0.140244	0.124060	0.119353	0.157199
TOT.		499,640				
No Funcional	NF	3,485,559	0.143736	0.133256	0.132627	0.154290
No Funcional conserv.	NFc	1,329,548	0.142306	0.132223	0.131684	0.152756
TOT.		3,485,559				

Tabla 4*. La tabla muestra el número de SNPs catalogados según distintas categorías de funcionalidad. En total se han encontrado 499,640 SNPs en elementos reguladores situados en genes humanos y/o en sus zonas 5Kb río arriba. En esas mismas zonas se buscaron el resto de SNPs que no se han predicho como funcionales porque, o bien no caen en ningún elemento regulador, o bien sus alelos mutados no producen un cambio significativo en el score predicho con el alelo original y por tanto no se considera que puedan tener un efecto funcional. En total se encontraron 3,485,559 SNPs no funcionales. Para cada categoría se anotó la media de las frecuencia del alelo minoritario (MAF) de sus SNPs en las cuatro poblaciones de HapMap.*

Según estos datos podemos ver que los SNPs con potencial efecto funcional presentan, en general, una frecuencia alélica significativamente menor que los SNPs que tomamos como controles (figura 17). Además si comparamos para cada categoría, la frecuencia es generalmente menor en SNPs conservados que en SNPs no conservados, y este patrón es consistente en las cuatro poblaciones de HapMap.

Como se puede observar la mayoría de estos SNPs son variantes comunes (MAF > 5%). Esto es algo que se puede esperar ya que, aunque una proporción de variantes raras se han incluido explícitamente debido a su función o localización, la mayoría de las variantes genotipadas en HapMap son variantes comunes. Sin embargo los SNPs situados en sitios de *splicing* muestran una MAF media mucho menor que el resto, algo que podría estar indicando que estos son los SNPs que potencialmente podría ser más dañinos.

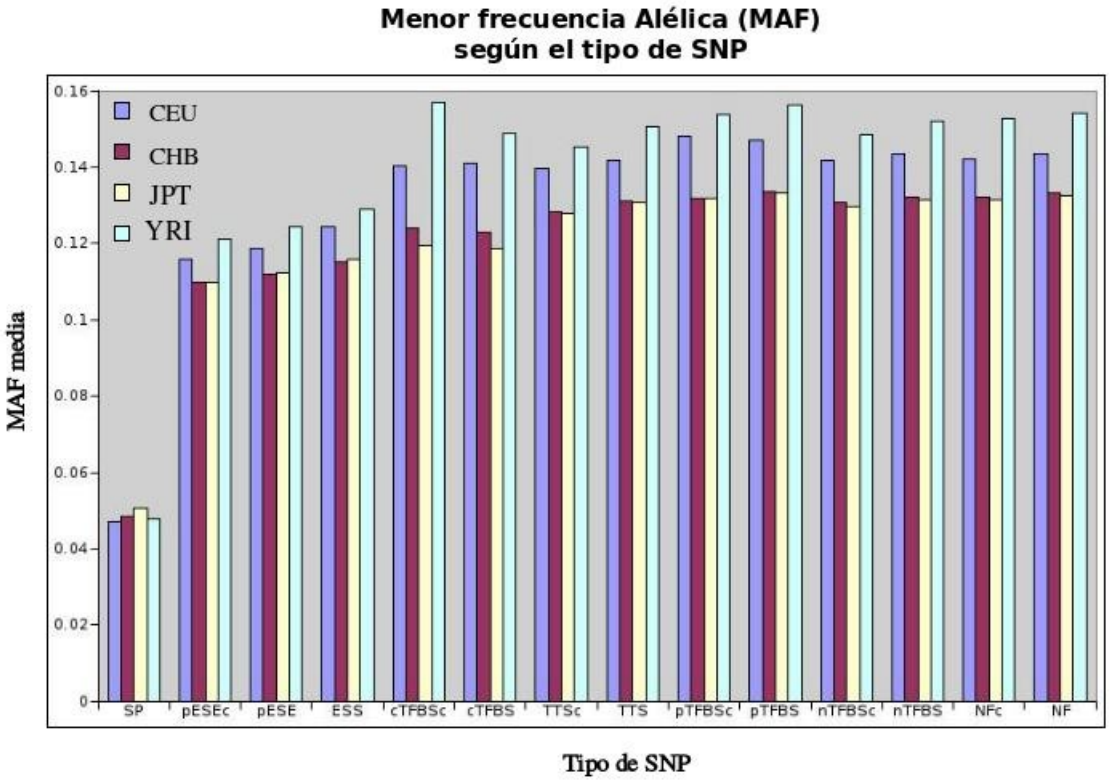


***Figura 17**. Representación de las frecuencias del alelo minoritario (MAF) para cada categoría funcional en las cuatro poblaciones de HapMap. La figura muestra que en general la MAF es menor en las categorías funcionales que en las no funcionales (NFc y NF), siendo los SNPs situados en sitios de splicing (SP) los que tienen una MAF significativamente más pequeña. En la figura también se aprecia que las MAF son mayores en población africana (YRI) que en el resto de poblaciones.*

Para cada categoría también se buscó el porcentaje de SNPs que esta presentes en sólo una de las poblaciones de HapMap, y se encontró que los SNPs localizados en sitios de *splicing*, potenciadores, silenciadores, TTSs y algunas categorías de SNPs en sitios de unión a factor de transcripción tienen una tendencia a ser más específicos de población que los controles (figura 18). Por el contrario, estos mismos SNPs tiene una tendencia menor a aparecer en las cuatro poblaciones de HapMap (figura 19).

Una excepción a esta regla son los SNPs en los que se ha predicho una pérdida o aparición de un nuevo TFBS, ya que aparecen con una MAF media similar a la de los etiquetados como no funcionales, y parecen ser ligeramente menos específicos y más comunes a todas las poblaciones. Quizá esto podría deberse a que el método empleado en la predicción de SNPs en TFBSs presenta un elevado número de falsos positivos, y puede estar indicando que este método, utilizando la base de datos de TRANSFAC® puede no ser el mejor criterio para la selección de SNPs, algo que ya se ha sugerido en otras publicaciones (Montgomery et al., 2007).

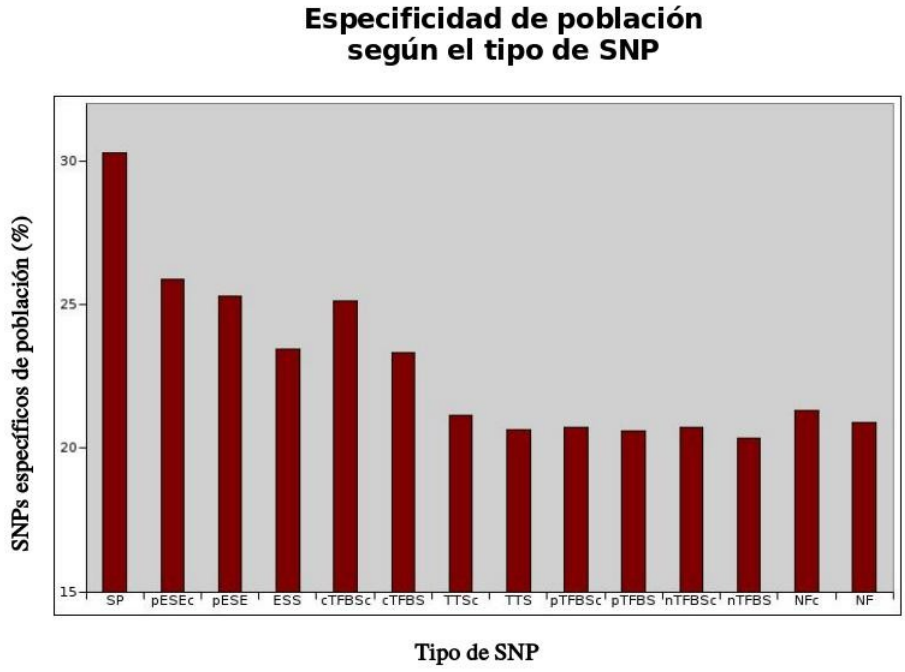


Figura 18. *Para cada categoría funcional se representa el porcentaje de SNPs que aparecen en una sola población de HapMap (CEU, CHB, JPT o YRI). Los SNPs situados en sitios de splicing (SP), ESEs, ESSs y los SNPs cuyo alelo mutante resulta en un cambio en el score que predice la presencia de un TFBS (cTFBS, cTFBSc), son más específicos que los no funcionales (NF, NFc).*

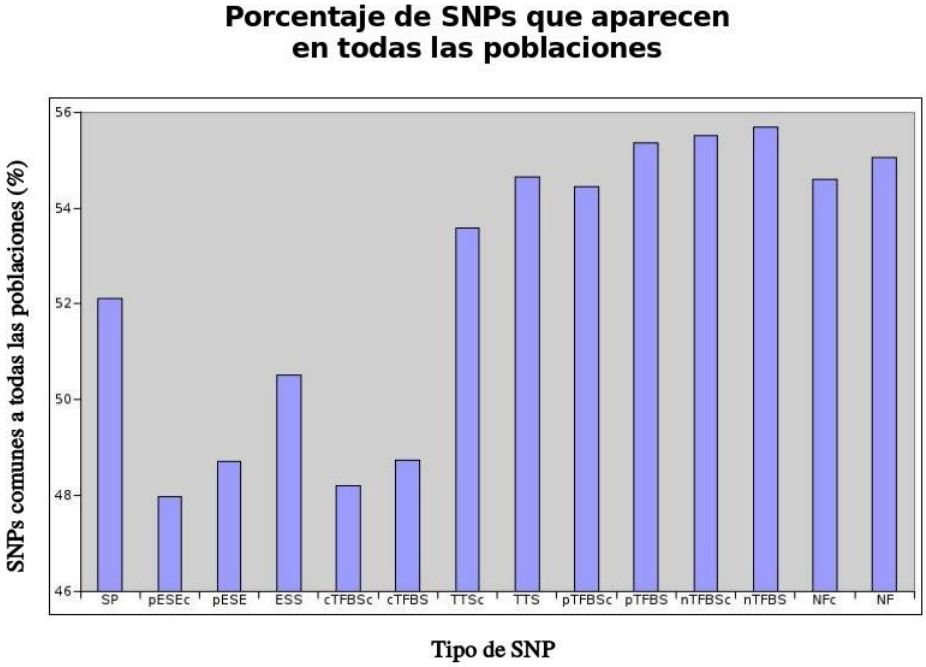


Figura 19. *Para cada categoría funcional se representa el porcentaje de SNPs que aparecen en todas las poblaciones de HapMap (CEU, CHB, JPT y YRI). La figura muestra que en general los SNPs catalogados como funcionales son menos comunes a todas las poblaciones de HapMap.*

Estos resultados son consistentes con el efecto de la selección purificadora de reducir la frecuencia de alelos dañinos, y aunque el hecho de que un SNP tenga una frecuencia alélica baja y sea especifico de población no implica que sea funcional, los resultados sugieren que el testar frecuencias alélica en la población de interés puede ser otro efecto a tener en cuenta a la hora de seleccionar SNPs para estudios de asociación genéticos.

1.7. SNPs codificantes no-sinónimos (nsSNPs)

1.7.1. Presión selectiva en nsSNPs

El efecto de la selección purificadora, o presión selectiva en un término más amplio, ha sido largamente estudiado en nsSNP (Hughes *et al*., 2003). La presión selectiva puede ser utilizada, a través de métodos evolutivos que tengan en cuenta información filogenética, como método para testar desviaciones de la neutralidad en aquellos codones que contengan SNPs, y por tanto para predecir la posible patogenicidad de nsSNPs (Arbiza *et al*., 2006).

La presión selectiva puede medirse como ω=dN/dS, donde dN y dS son las tasas de mutaciones no sinónimas y sinónimas respectivamente. Arbiza y colaboradores utilizan dos aproximaciones, el modelo de máxima verosimilitud (Yang y Nielsen, 2002) implementado en el paquete PAML (Yang, 1997), y el método de *likelihood-ratio* (SLR) (Massingham y Goldman, 2005), para estimar los valores de ω. De acuerdo con los autores, los codones con mutaciones que se encuentran frecuentemente asociadas a enfermedad tienen valores de ω menores de 0.1 y por tanto nsSNP localizados en esos codones tienen alta probabilidad de ser patológicos.

Para predecir el posible efecto funcional de los nsSNPs por medio del método descrito arriba, obtuvimos de la base de datos Ensembl todos los nsSNPs y anotamos sus cambios de aminoácido, su posición en la secuencia proteica y la secuencia aminoacídica flanqueante. Con estos datos se estimó la presión selectiva utilizando las secuencias ortólogas de mamíferos disponibles en Ensembl.

De los 60,592 nsSNPs descritos en la versión 39 de Ensembl, sólo se encontraron predicciones de ω para 16,983 (un 28%). Esta cifra es aún más baja si en lugar de alineamientos con secuencias de mamíferos se utilizan secuencias de vertebrados, al incluirse más secuencias. El porcentaje tan bajo de SNPs con predicciones se debe a que los SNPs mapean zonas donde no hay un número suficiente de

secuencias ortólogas para alinear o a que mapean en huecos en los alineamientos.

De los 16,983 SNPs con predicciones, hay 8,380 cuyos estadísticos por cualquiera de los dos métodos (PAML o SRL) son menores de 0.1, y de ellos, 6,609 son SNPs cuyos valores de estadístico son menores de 0.1 en ambos métodos. Es decir, aproximadamente un 13% de los nsSNPs tienen un potencial efecto patológico teniendo en cuenta la estima de la presión selectiva a nivel de codón.

Es importante saber que el método no considera el efecto que una mutación pueda tener sobre sitios con funcionalidades añadidas como sitios que afectan al *splicing* o SNPs que generan codones de terminación. Al no poder integrarse esta información, las predicciones para esos codones podrían ser falsos positivos.

Para los 8,380 SNPs anotados como patológicos por cualquiera de los dos métodos de estima de presión selectiva, se buscaron también los datos de frecuencias alélica en las poblaciones de HapMap. Estos datos se compararon con las frecuencias del resto de nsSNPs para los que se ha podido hacer predicción y no se etiquetaron como patológicos por ninguno de los métodos (tabla 5, figura 20).

Tipo SNP	*Código*	*Número SNPS (% del total de nsSNPs)*	*media MAF CEU*	*media MAF CHB*	*media MAF JPT*	*media MAF JPT*
nsSNP patológicos (PAML o SRL)	nsSNPp	8,380 (13.83%)	0.06124	0.06251	0.0634	0.05919
nsSNP no predichos como patológicos	nsSNPnp	8,603 (14.19%)	0.1277	0.11973	0.11967	0.13051
TOT.		16,983 (28.02%)				

***Tabla 5**. La tabla muestra el número de SNPs catalogados como patológicos (nsSNPp) o no patológicos (nsSNPnp) según la estima de presión selectiva. Solo se pudieron hacer predicciones para un 28.02% del total de nsSNPs, y aproximadamente la mitad de ellos resultaron ser patológicos por cualquiera de los dos métodos PAML o SRL. Para las dos categorías (patológicos o no) se anotó la media de las frecuencias del alelo minoritario (MAF) de sus SNPs en las cuatro poblaciones de HapMap.*

En este caso también se puede observar que los SNPs catalogados como patológicos tiene una MAF significativamente menor y son más específicos de población que los que no se predijeron como patológicos (figura 20).

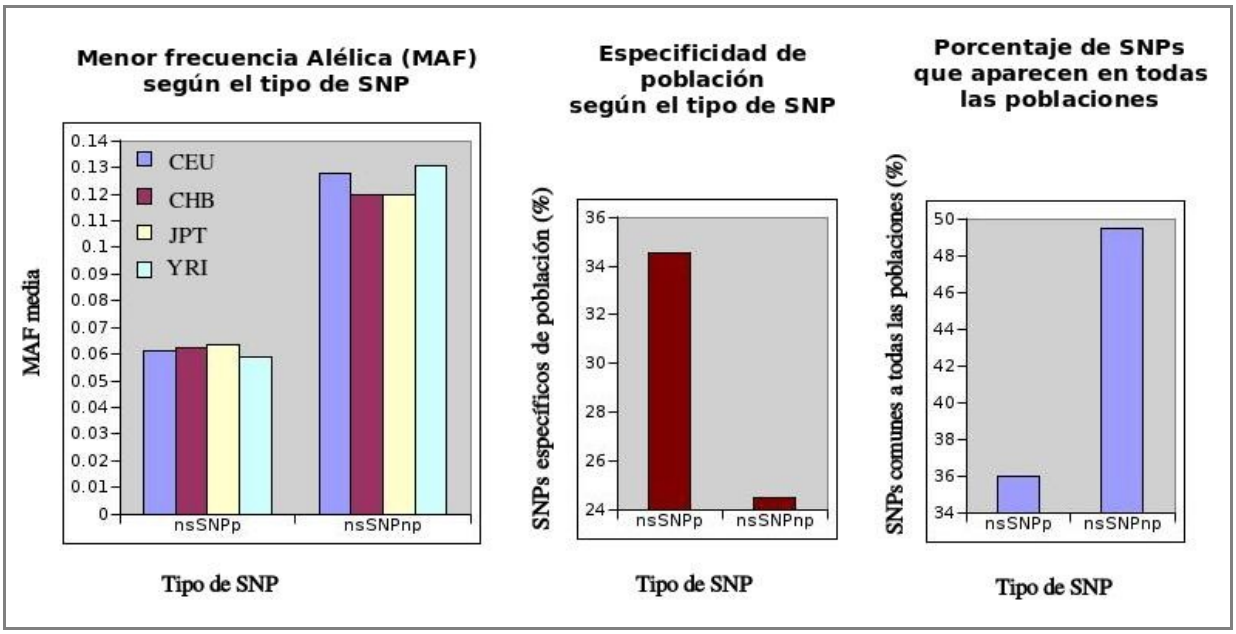


Figura 20. *En la figura se muestra, para SNPs patológicos (nsSNPp) y no patológicos (nsSNPnp), las frecuencias del alelo minoritario (MAF) en las cuatro poblaciones de HapMap* ***(izda)****, el porcentaje de SNPs que aparecen en una única población* ***(centro)*** *y el porcentaje de SNPs que aparecen en las cuatro poblaciones* ***(dcha)****. Los datos muestran que los SNPs predichos como patológicos por estima de presión selectiva tienen MAFs menores y son más específicos de población que los no patológicos.*

2. Herramientas bioinformáticas para la selección de SNPs: PupaSNP, PupasView y PupaSuite

Con el objetivo de facilitar el entendimiento de las implicaciones funcionales de los polimorfismos identificados en estudios de asociación, durante esta tesis se han desarrollado un conjunto de herramientas para el análisis de SNPs. Estas herramientas integran las anotaciones de varias bases de datos y generan, a partir de los métodos bioinformáticos descritos en la sección anterior, información funcional para todos los polimorfismos del genoma humano.

Esta información precalculada está almacenada en una base de datos que es públicamente accesible a través de las herramientas, con el objetivo de aportar a la comunidad científica un mecanismo para aumentar la eficacia en la exploración de la función de SNPs. Estas herramientas se prevé que estén continuamente mejorando por laadición de nuevas funcionalidades y anotaciones.

Desde la primera publicación en 2004, estas herramientas se han citado en 33 trabajos científicos,

y la media de unos 100 accesos diarios muestra que son herramientas con gran aceptación en el ámbito científico.

2.1. PupaSNP

PupaSNP (Conde *et al*., 2004) fue la primera herramienta en desarrollarse y es la que ha dado el nombre a la saga. PupaSNP se diseñó como herramienta de alto rendimiento para la búsqueda de SNPs.

La herramienta recoge una lista de genes y genera un informe donde se listan todos los SNPs con potencial efecto funcional de esos genes y de sus regiones promotoras. Los genes se pueden seleccionar directamente por su localización en una región cromosómica (especificando citobandas o posiciones genómicas), o pueden introducirse como una lista de genes, tanto genes no relacionados como genes de una determinada ruta metabólica o implicados en alguna función común.

Para esos genes y sus regiones promotoras PupaSNP encuentra todos los SNPs que podrían causar una pérdida o alteración de la funcionalidad. Proporciona además información funcional para los genes obtenida de bases de datos como OMIM y Gene Ontology, así como información sobre genes homólogos en otras especies. De esta forma, la consideración de los SNPs en un contexto funcional puede servir de ayuda para entender las implicaciones biológicas potenciales de los SNPs y genes estudiados.

Además la herramienta incorpora una opción en la que el usuario puede hacer predicciones sobre SNPs que no estén incluidos en las bases de datos. Indicando la posición del SNP con respecto al gen más próximo, y sus alelos, la herramienta localiza si el SNP esta situado en algún motivo regulador importante como en TFBSs, sitios de *splicing* o ESEs y predice el efecto que puede tener el cambio de alelo.

En la figura 21 se muestra un ejemplo de datos obtenidos con PupaSNP.

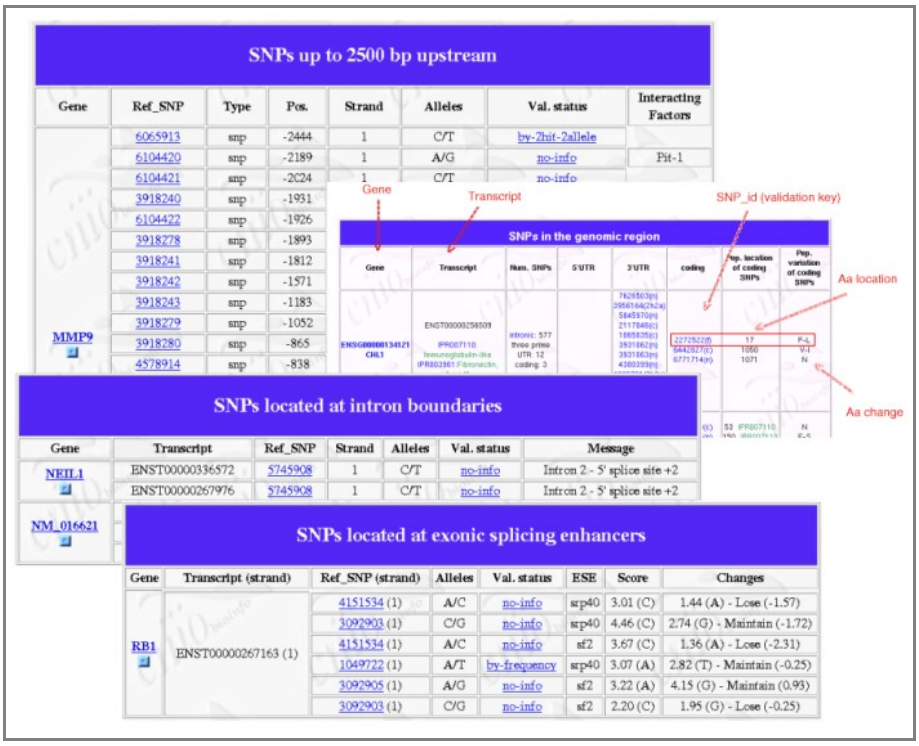


***Figura 21**. Selección de resultados de PupaSNP. Se muestran, para una lista de genes, sus SNPs localizados en TFBS, los nsSNPs, SNPs en sitios de splicing y SNPs en ESEs.*

2.2. PupasView

Además de la información funcional, la información sobre datos de frecuencias alélicas en distintas poblaciones es otro factor importante que debe tenerse en cuenta a la hora de seleccionar SNPs. Así, polimorfismos infrecuentes pueden ser de poco interés como marcadores. Además, el desequilibrio de ligamiento (LD) es otro factor interesante a la hora de seleccionar SNPs ya que si dos SNPs están en LD uno solo de ellos proporcionará la información suficiente para los análisis de ligamiento o asociación.

Con la idea de añadir esta información se desarrolló en 2005 la segunda herramienta llamada PupasView (Conde *et al.*, 2005). PupasView puede usarse tanto sola como acoplada a PupaSNP. Mientras que la primera herramienta se centra en la selección de SNPs, PupasView se diseñó, además de como herramienta de selección, como herramienta de visualización gráfica y se centra en un solo gen.

PupasView funciona como un selector donde diferentes filtros basados en funcionalidad y frecuencias poblacionales pueden ser aplicados interactivamente sobre parámetros de LD con el

objetivo de obtener una selección óptima con el mínimo número de SNPs que contengan la máxima información en la región del gen de interés.

PupasView recoge como parámetro de entrada el nombre de un gen (introducido con cualquier tipo de identificador aceptado en Ensembl) y la longitud de la región que flanquea al gen y en la que se quieran buscar TFBSs.

Si se ejecuta PupasView con los parámetros que aparecen por defecto, se obtiene una imagen similar a la figura 22, donde todos los tránscritos y SNPs del gen y de la región flanqueante están representados. Esto normalmente supone un número muy grande de SNPs, y para seleccionar un subgrupo de ellos que sean más informativos la herramienta proporciona diferentes filtros:

- Estado de validación, que es un parámetro que da información sobre la calidad del SNP, ya que indica si el SNP se ha observado en múltiples e independientes fuentes, si se ha descubierto por métodos computacionales o por resecuenciación, etc.
- Tipo de SNP (codificante, UTR...), que hace referencia a la posición del SNPs en el gen.
- Frecuencia y población, que incluye la posibilidad de filtrar por un rango de frecuencias del alelo menos frecuente en una o más poblaciones.
- Propiedades funcionales, incluyendo SNPs en TFBSs, en ESEs, en sitios de *splicing* y en TTSs (todos o sólo los localizados además en regiones conservadas) o nsSNPs predichos como patológicos con el programa Pmut.

Para el cálculo de los bloques y parámetros de LD en la herramienta PupasView, primero se recogen los SNPs seleccionados por el usuario según los criterios de funcionalidad, validación y frecuencia poblacional elegidos. Para aquellos SNPs con datos de genotipado en HapMap, la herramienta muestra los datos de LD entre SNPs contiguos y los bloques de haplotipos generados con el programa Haploview implementado en PupasView (figura 22).

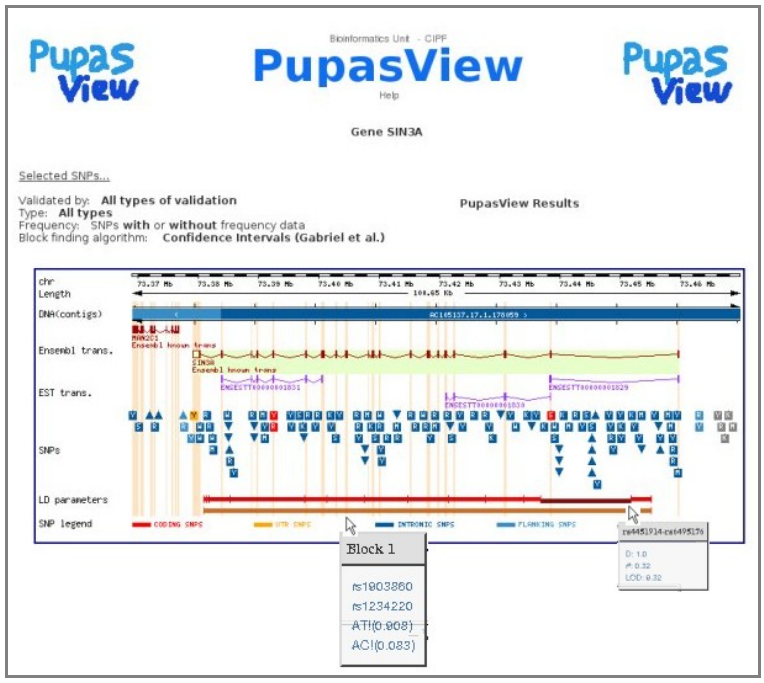


***Figura 22**. Resultados de PupasView. La figura muestra todos los SNPs del gen SIN3A, representados en cajas coloreadas según el tipo de SNP (codificante, UTR...), así como el gen (resaltado en amarillo) y los tránscritos de los alrededores. Los valores de LD entre dos SNPs contiguos genotipados en HapMap son mostrados gráficamente mediante rectángulos coloreados que van desde un color más claro (r^2 bajo) a uno más oscuro (r^2 alto), donde el color es azul, si LOD < 2, o rojo, si LOD >= 2. Los bloques de haplotipos son mostrados con rectángulos de color marrón, que se extienden desde el primer al último SNP del bloque. Al pasar el cursor sobre los rectángulos, aparece un texto en el que se muestran los SNPs y los haplotipos (con las frecuencias de HapMap entre paréntesis) de cada bloque. Los Tag SNPs aparecen señalados con un una marca (!).*

2.3. PupaSuite

Posteriormente, y para integrar PupaSNP y PupasView en un único paquete de programas integrado, se creó PupaSuite (Conde *et al.*, 2006). En PupaSuite no sólo se mejoró la funcionalidad de las herramientas sino que se implementaron nuevas facilidades como el análisis de datos de genotipado del usuario para derivar haplotipos con información funcional o la inclusión de nuevas predicciones para el análisis de SNPs no-sinónimos.

Aunque las tres herramientas coexistieron durante un año, desde 2007 las dos primeras están redirigidas a PupaSuite. Siguiendo la filosofía de PupaSNP, PupaSuite permite introducir tanto una lista de genes como una región cromosómica, lo que corresponde con los dos tipos de análisis más comunes: genes relacionados con una enfermedad porque están funcionalmente relacionados (por ejemplo, pertenecen a una ruta afectada en la enfermedad), o genes presentes en una región cromosómica ligada a la enfermedad. En ambos casos la herramienta devuelve una lista de SNPs con

sus putativos efectos funcionales, y en el caso de regiones cromosómicas también es posible buscar bloques de haplotipos.

PupaSuite también puede analizar directamente listas de SNPs. En este caso además del putativo efecto funcional también es posible obtener información sobre las frecuencias alélicas en diferentes poblaciones, según las anotaciones de Ensembl, así como haplotipos y tags para hacer una preselección de SNPs para genotipado. De una lista de SNPs también se puede obtener información sobre LD. Supongamos que hemos realizado un estudio en el que hemos encontrado una serie de SNPs asociados con un haplotipo de riesgo. Podemos introducir esta lista en la herramienta y sacar todos los SNPs genotipados en HapMap que estén en LD con nuestro grupo de SNPs. De esta forma podemos hacer un análisis funcional de todos los SNPs en LD con el conjunto de SNPs originales para identificar putativos SNPs causativos.

Una opción nueva en PupaSuite es el análisis de haplotipos funcionales. Esta opción permite al usuario testear sus datos de genotipado para encontrar haplotipos con SNPs funcionales. En este paso se pueden analizar datos de estudios de casos y controles en los que se puede ver diferencias en las frecuencias alélicas en SNPs funcionales entre los dos grupos (figura 23).

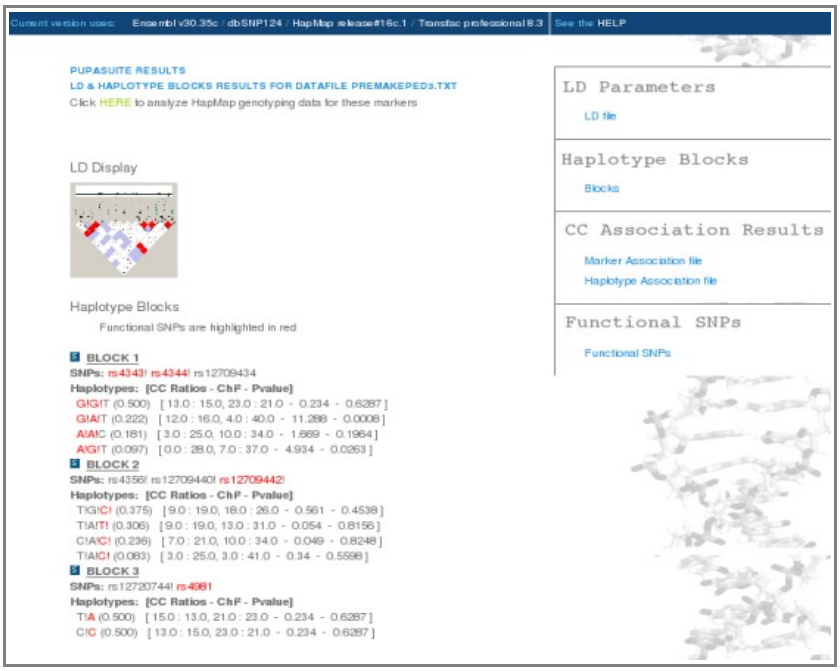


Figura 23. *Resultados de PupaSuite. La figura muestra la imagen de LD para los SNPs analizados, los cuales aparecen en rojo si están localizados en TFBSs, ESEs, ESSs, sitios de splicing, TTSs o si son nsSNPs catalogados como patológicos por el algoritmo de Pmut, la base de datos SNPeffect o por estima de presión selectiva.*

Además de analizar las propiedades funcionales incluidas en las anteriores herramientas también se han incluido análisis de polimorfismos en ESSs y se prevé la introducción de métodos adicionales para la predicción de SNPs en TFBSs y sitios de *splicing*. Con respecto al putativo impacto de nsSNPs, en el momento de escribir esta tesis la herramienta incorpora predicciones obtenidas por el programa Pmut y por estima de presión selectiva. Además, la base de datos de PupaSuite incluye las anotaciones de la base de datos SNPeffect. En un esfuerzo conjunto las dos bases de datos se han sincronizado para proporcionar anotaciones para SNPs codificantes y no codificantes en una sola base de datos y de esta forma proporcionar una información valiosa para interpretar y guiar experimentos (figura 24).

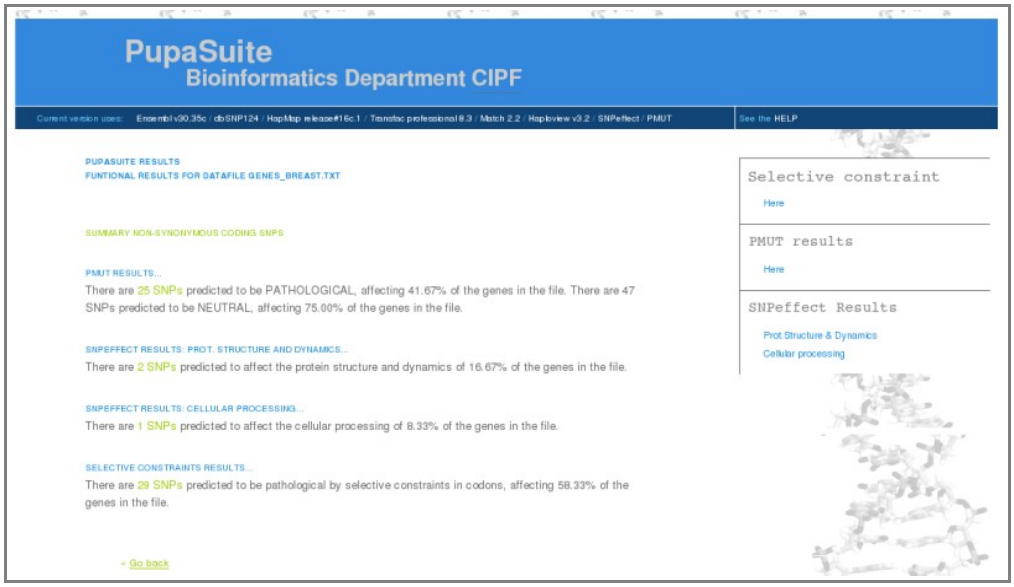


Figura 24. *Resultados de PupaSuite. La figura muestra los resultados obtenidos del análisis de los nsSNPs de una lista de genes por los tres métodos incorporados en PupaSuite. Un enlace en la parte superior de la imagen muestra un resumen con las predicciones de los tres métodos.*

3. Análisis de variaciones de número de copia: ISACGH

El descubrimiento de las variaciones de número de copia (CNVs) como una característica universal en los genomas, ha coincidido con el interés creciente en el estudio de la influencia de la variación genómica en enfermedades y evolución. Los SNPs son actualmente el tipo de variación genómica más

estudiada debido a su estabilidad y a su abundancia en el genoma. Sin embargo, algunos estudios indican que el contenido en CNVs en humanos incluso podría llegar a exceder el de SNPs (Lee, 2005), por lo que probablemente en el futuro los estudios de asociación de casos y controles de genomas enteros empezarán a incorporar análisis de CNVs. Mientras tanto, la aproximación mas común para el estudio de CNVs son los *arrays* de hibridación genómica comparativa (aCGH).

Con el objetivo de proporcionar a la comunidad científica una herramienta para el estudio de este segundo tipo de variación genómica, los CNVs, durante esta tesis se ha desarrollado una herramienta para el análisis de aCGHs llamada ISACGH (Conde *et al.*, 2007a, 2007b).

El programa

ISACGH es una herramienta web que permite el análisis combinado de alteraciones en el número de copia y expresión génica. La herramienta recoge una lista de genes con sus datos de expresión (a partir de *microarrays* de mRNA), sus valores de hibridación genómica (a partir de aCGHs) o ambos a la vez, y mapea esos valores en el genoma (humano o ratón) de forma gráfica. A partir de los valores de hibridación genómica, el programa predice las regiones con alteraciones de número de copia a través de 4 métodos distintos (ver material y métodos).

La representación conjunta de los dos tipos de datos (expresión y genómicos) permite una primera evaluación visual del efecto de las CNVs en la expresión global de los genes contenidos en la región delecionada o amplificada. Además el programa incorpora un test de la t que permite evaluar expresión diferencial entre los genes con número de copia normal y los genes situados en las regiones con alteraciones.

La herramienta proporciona diferentes posibilidades para la representación de los resultados dependiendo del foco de estudio, ya que se pueden representar todos los cromosomas de una muestra o un solo cromosoma para múltiples muestras (figura 25). Así, la representación del cariotipo completo de una muestra puede ser útil en los análisis de genomas enteros, mientras que la representación a nivel de cromosoma es apropiado para detectar CNVs de loci relativas al resto de loci del mismo cromosoma, con independencia de la ploidía.

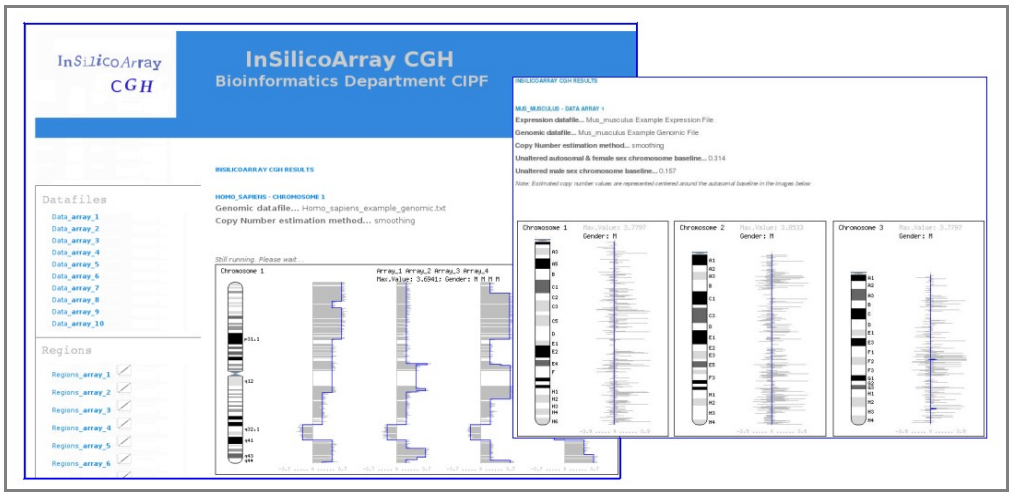


Figura 25. *Resultados de PupaSuite. La figura muestra los resultados obtenidos del análisis de los nsSNPs de una lista de genes por los tres métodos incorporados en PupaSuite. Un enlace en la parte superior de la imagen muestra un resumen con las predicciones de los tres métodos.*

Integración de anotaciones de Ensembl, información funcional y el paquete GEPAS

La herramienta incorpora un zoom interno para ver los resultados en detalle, y además permite representarlos en el navegador de Ensembl a través de su sistema de anotación distribuida (DAS).

El DAS es un sistema cliente-servidor donde un cliente, en este caso Ensembl, integra información de muchos servidores (ver http://www.biodas.org). Utilizando la arquitectura DAS, Ensembl recoge información sobre anotaciones genómicas de muchos sitios web, integra esa información y la muestra al usuario junto con sus propias anotaciones y datos. De esta forma el uso de servidores DAS para la visualización de cualquier característica genómica en el visor de Ensembl proporciona un escenario excelente para el estudio de los resultados producidos por ISACGH en un contexto genómico, con la posibilidad de acceder a cualquier tipo de información disponible en Ensembl (figura 26).

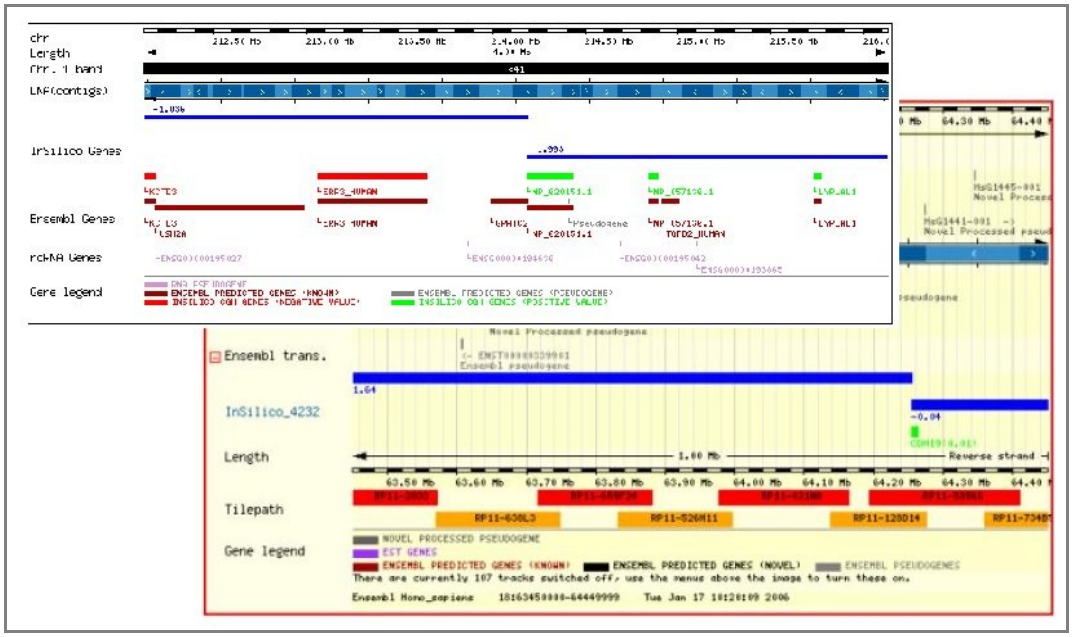


Figura 26. *Resultados de PupaSuite. La figura muestra los resultados obtenidos del análisis de los nsSNPs de una lista de genes por los tres métodos incorporados en PupaSuite. Un enlace en la parte superior de la imagen muestra un resumen con las predicciones de los tres métodos.*

Además de la información procedente de Ensembl, ISACGH incorpora información funcional obtenida a través de la herramienta FatiGO (Al-shahrour *et al.*, 2004), que emplea el test exacto de Fisher para determinar el enriquecimiento en diferentes categorías funcionales (rutas de KEGG, motivos de InterPro, Gene Ontology, etc.) entre los genes localizados en las regiones delecionadas/amplificadas detectadas y el resto de genes del cromosoma.

Por ultimo, aunque ISACGH es una herramienta independiente, está completamente integrada en el paquete GEPAS (Herrero, *et al.*, 2003; Montaner *et al.*, 2006). GEPAS es un servidor web que incorpora los principales métodos para el análisis de *microarrays*, y la integración del ISACGH en GEPAS le proporciona ventajas adicionales como la posibilidad de realizar normalizaciones o preprocesar los *microarrays* antes de su utilización con ISACGH. Por ejemplo, después de detectar un grupo de genes que coexpresan o que correlacionan con un rasgo determinado, puede ser muy interesante ver donde mapean en el genoma.

4. Análisis de datos de genotipado

Una vez seleccionados el conjunto óptimo de SNPs para un estudio de asociación el siguiente paso, después de obtener los datos de genotipado, es el análisis e interpretación de esos datos.

En este sentido, la parte final de esta tesis se ha centrado en el desarrollo de un programa que permite analizar datos de genotipado en combinación con información biológica en el contexto de estudios de asociación.

En un primer paso se seleccionan un conjunto de marcadores mediante el cálculo de estadísticos Chi cuadrado que miden la asociación alélica de cada SNP con la enfermedad. Estos estadísticos individuales sirven para ordenar los marcadores según diferencien mejor entre casos y controles. A partir de esos estadísticos locales y la información biológica procedente de bases de datos de interacciones proteína-proteína (test PP), Gene Ontology (test GO) o conservación de secuencia (test C), se genera un estadístico global o *score*.

La significación estadística de ese *score* se mide permutando repetidamente las etiquetas de casos y controles y recalculando los estadísticos locales y global para obtener una distribución del *scores* bajo la hipótesis nula de no asociación. La proporción de *scores* permutados que exceden el *score* visto en los datos originales se aproxima al P valor de ese *score*. Si este P valor obtenido por permutaciones resulta ser significativo se puede usar para derivar una nueva hipótesis en el mecanismo de la enfermedades estudiada.

4.1. Aplicación

El método se ha aplicado en un estudio de asociación en el que se han genotipado 116,204 SNPs utilizando el *microarray* GeneChip Mapping 100K de Affymetrix en un panel de 184 individuos, 96 de ellos pacientes con asma y los 88 restantes controles.

4.1.1. Análisis preliminar

Debido a la diferente distribución de varones y hembras en casos y controles, se decidió eliminar los SNPs localizados en el cromosoma X para reducir la probabilidad de falsas asociaciones. También

se eliminaron SNPs con más de un 10% de datos incompletos, ya que, aunque en el análisis de SNPs individuales la pérdida de algunos genotipos no es muy importante, en análisis de muchos marcadores puede ser problemático (Balding, 2006).

El desequilibrio de Hardy-Weinberg puede ser utilizado como control de calidad, ya que valores extremadamente altos podrían estar indicando errores de genotipado. Por ese motivo también se eliminaron aquellos SNPs donde la distribución alélica en la población control no satisface el equilibrio de Hardy-Weinberg para un nivel de significancia α=0.01 ($\chi^2 > 6.635$, α=0.01, 1df).

Al final más de 100,000 SNPs se seleccionaron para el estudio.

4.1.2. Test de P valores (test PV).

La suma de estadísticos individuales se ha propuesto como primer método de análisis en estudios de casos y controles con múltiples SNPs. En esta aproximación se selecciona un conjunto de marcadores, se combina la contribución de cada uno de ellos mediante la suma de sus estadísticos individuales y se testan contra la hipótesis nula de que ninguno de ellos está asociado con la enfermedad.

En el primer paso se evalúa la asociación alélica de cada SNP individual mediante el test de χ^2. En un segundo paso los SNPs se ordenan de forma creciente por su χ^2 y se calcula un *score* para los primeros 'n' SNPs tal y como se describe en la sección de material y métodos. El parámetro n puede modificarse; un número demasiado bajo de SNPs puede ser poco informativo, y la elección de demasiados puede añadir mucho ruido al análisis. En este caso hicimos el corte en 100 SNPs (Pval < 0.01) ya que parece un número razonable de SNPs para nuestros datos.

En un tercer paso se determina el nivel de significación mediante un test de permutación realizado bajo la hipótesis nula de no asociación. En este punto el testeo múltiple de todos los SNPs se elimina al realizarse un único test.

Este test se aplicó a los datos y después de 10,000 permutaciones un 98.32% de los *scores* permutados S_p fueron mayores que el original S^* (p=0.0168) (figura 27). Es decir, solo 1 de cada 100 veces esperaríamos ver un *score* tan pequeño como el original bajo la hipótesis nula de no asociación. Esto nos podría estar indicando que las distribuciones alélicas están significativamente asociadas con la presencia de la enfermedad en estudio.

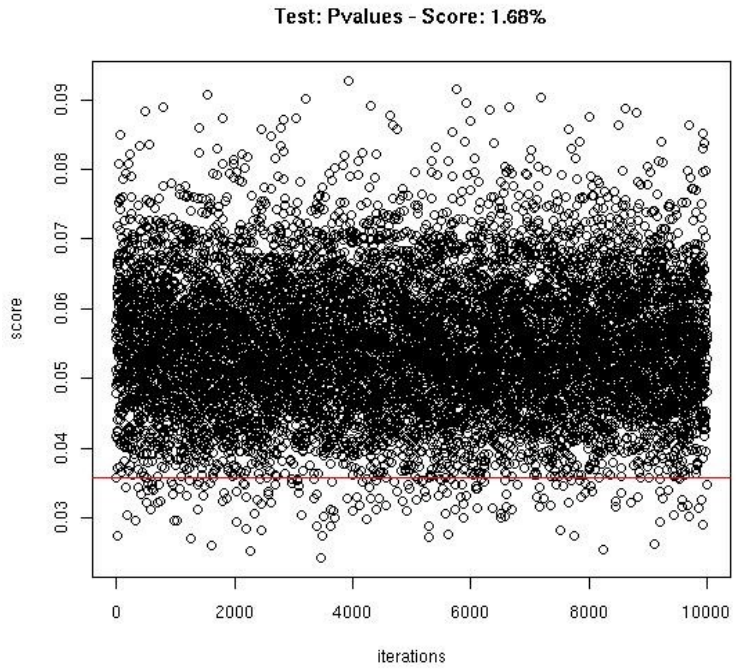


Figura 27. *La figura representa, mediante puntos, la distribución de scores obtenidos por permutación. La linea roja indica el score obtenido sin permutar (S*). Solo un 1.68% de los S_p obtenidos al azar son menores que S*.*

4.1.3. Test de Gene Ontology (test GO)

Los métodos *multi-marker* descritos en la introducción son métodos que, a la hora de detectar SNPs con asociación a enfermedad, tienen en cuenta que las enfermedades complejas pueden estar causadas por un patrón de variantes genéticas que interaccionan entre ellas, y que el riesgo asociado a un locus determinado puede estar influenciado por el genotipo de otro locus distinto.

En general, las interacciones de riesgo entre loci son más creíbles entre genes implicados en interacciones físicas o entre genes que participan en la misma ruta metabólica o en las mismas redes reguladoras (Carlson *et al.*, 2004). Por eso, es concebible pensar que el uso de estadísticos de grupo para SNPs relacionados funcionalmente, añadido a los estadísticos de SNPs individuales, puede incrementar la sensibilidad de detectar múltiples SNPs implicados en la enfermedad de estudio.

Partimos de la suposición de que grupos de SNPs pueden influir colectivamente en un proceso biológico común. De esta manera, podemos incluir la información biológica disponible en Gene Ontology para dar más peso o credibilidad al grupo de SNPs que mejor distinguen entre casos y controles si participan en algún proceso biológico común subyacente.

Como se ha descrito anteriormente, en el primer paso se seleccionan un grupo de genes asociados

a un grupo de SNPs. Para el conjunto de genes se calcula un *pairwise score S**, como la suma ponderada de los estadísticos locales, que tiene en cuenta no sólo los estadísticos de los SNPs sino también los procesos biológicos compartidos subyacentes, dando mayor peso a los pares que comparten términos que son más específicos. No es necesario corregir por testeo múltiple ya que se realiza un único test estadístico (un χ^2 combinado ponderado, un estadístico global construido como función de estadísticos locales) que se evalúa mediante permutaciones.

La interpretación biológica de los resultados puede examinarse *a posteriori* analizando qué genes y qué funciones (o términos GO) son los que aparecen asociados al grupo de SNPs para evaluar si los genes implicados dan un sentido biológico a la enfermedad en estudio.

Como en el test anterior, se calcularon los χ^2 para cada SNP, se ordenaron por orden decreciente y se seleccionaron los 100 primeros SNPs. Se aplicó el test de GO y se obtuvo un conjunto de 34 genes derivados del grupo de 100 SNPs. Después de 10,000 permutaciones, se encontró que en un 99.47% de las veces (p=0.0053) el *score* S* original fue mayor que los obtenidos por permutaciones aleatorias (figura 28). Por tanto tenemos una fuerte evidencia de que la selección de genes obtenida podría ser influyente en el proceso de la enfermedad.

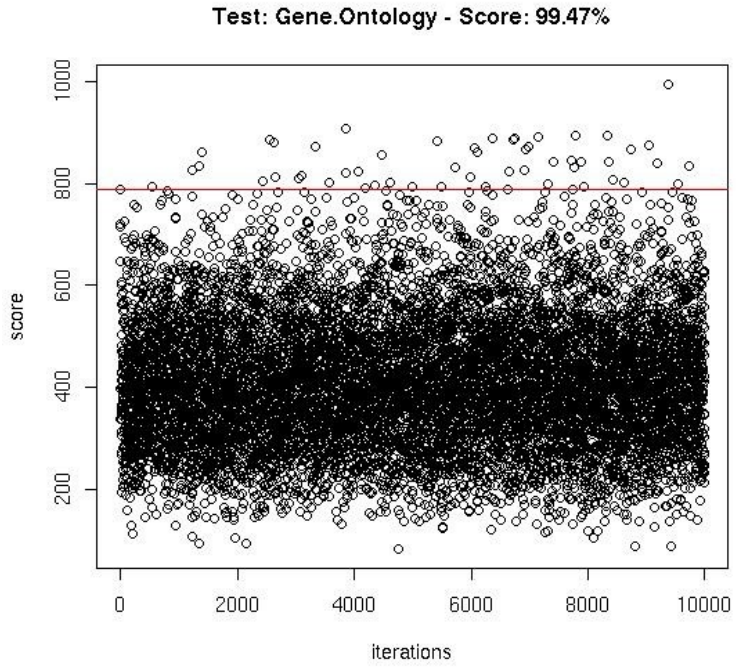


Figura 28. *La figura representa, mediante puntos, la distribución de scores obtenidos por permutación. La linea roja indica el score obtenido sin permutar (S*). Éste es mayor que el 99.47% de los S_p obtenidos por azar.*

4.1.4. Estratificación de poblaciones

Como ya se ha comentado, la asociación entre marcadores genéticos y enfermedad puede aparecer por causas distintas al ligamiento, como por ejemplo por estratificación de poblaciones, dando lugar a falsas asociaciones. Por tanto se testó si la asociación observada podría deberse a una posible estratificación y no a diferencias alélicas. Para ello se implementó el test de Pritchard-Rosenberg en el programa. Bajo la hipótesis nula de que las poblaciones tienen las mismas frecuencias alélicas, la suma del estadístico para todos los marcadores tiene una distribución de χ^2 con grados de libertad igual a la suma de grados de libertad de todos los marcadores. En la práctica, este estadístico evalúa la diferencia global de frecuencias alélicas entre casos y controles.

Se comprobó la estratificación de poblaciones en los datos originales utilizando 1,000 SNPs seleccionados al azar. Se calculó la suma de sus estadísticos χ^2 y se evaluó la significación estadística usando una distribución de χ^2 con 1,000 grados de libertad. El test mostró evidencia de estratificación en la población (χ^2 = 1114, 1000df, Pvalor*=6.700000e-03).

Corrección de la estratificación

Se seleccionaron un 5% de los SNPs uniformemente distribuidos a lo largo del genoma, y se utilizó el modelo de *admixture* implementado en el programa STRUCTURE2.1 (Pritchard *et al*., 2000a) con los siguientes parámetros: *i)* alpha=1, *ii)* iteraciones=20000 y *iii) burnin period*=20,000, para estimar la ascendencia poblacional de cada individuo. Con este programa se puede inferir la presencia de distintas poblaciones y asignar los individuos a una de las K poblaciones definidas *a priori*. Usando información anterior sobre la naturaleza poblacional de los individuos analizados, asumimos un número de poblaciones K=3.

Las proporciones ancestrales globales en cada población o grupo resultaron ser 0.219%, 0.555% y 0.227%, siendo las proporciones medias en los casos 0.199%, 0.597% y 0.204%, y en los controles 0.240%, 0.508% y 0.252%. Las estimaciones para cada individuo se muestran el la figura 29.

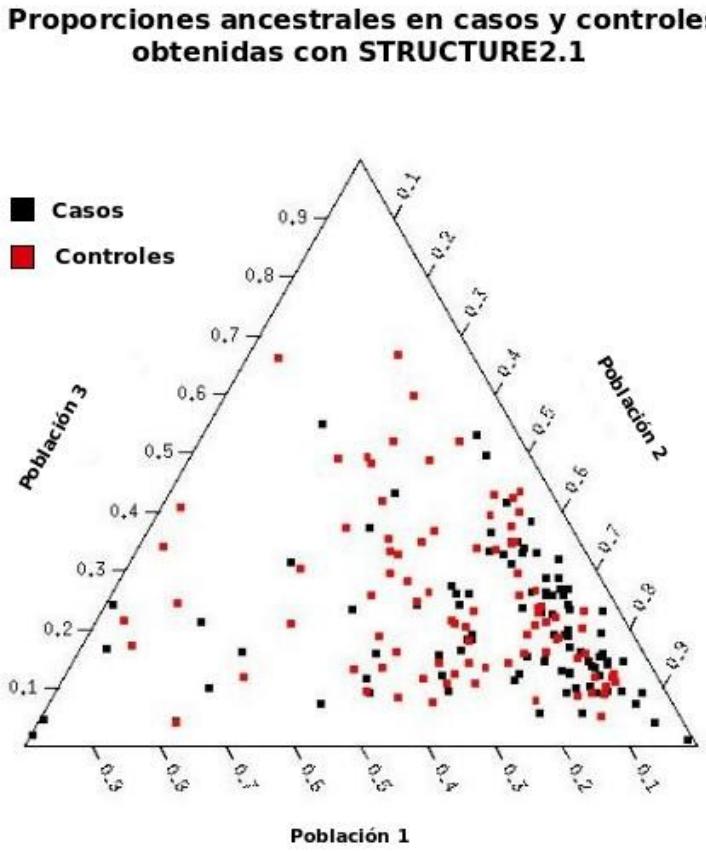


Figura 29. *Resultado gráfico obtenido con STRUCTURE2.1 asumiendo un número de poblaciones K=3. Cada vértice del triangulo representa una población parental. Cada punto identifica un individuo (casos en negro, controles en rojo) y su posición relativa en el diagrama corresponde al porcentaje ancestral de cada individuo en las 3 poblaciones.*

Se utilizaron los resultados obtenidos con STRUCTURE para aplicar el método PSAT (Kimmel *et al*., 2007) y realizar las permutaciones teniendo en cuenta la estructura poblacional. Los porcentajes ancestrales estimados por STRUCTURE se utilizaron para inferir vectores de probabilidad para cada individuo. De esta forma no se considera la hipótesis nula de que los individuos tiene igual probabilidad de tener enfermedad, sino que tiene mayor o menor probabilidad de acuerdo con el correspondiente componente de su vector de probabilidad.

Se repitió el test PV, pero en vez de permutar aleatoriamente las etiquetas de casos y controles se utilizó el modelo anterior (es decir, a cada permutación se le da un peso diferente de acuerdo con la estructura poblacional) para corregir la estratificación y se recalculó el porcentaje de veces que los *scores* S_p obtenidos con esos muestreos fueron mayores que el obtenido con los datos originales, y este porcentaje resultó ser un 97.67% (p=0.0233), ligeramente menor que antes (figura 30izda). Es decir, el

efecto de corregir la estratificación ha hecho que el P valor que mide la asociación sea más alto, aunque todavía se encuentra asociación de los mejores 100 SNPs con la enfermedad.

Además se repitió el análisis de GO utilizando los muestreos para corregir la estratificación. En este caso el porcentaje obtenido después de 10,000 iteraciones fue 99.54% (p=0.0046), indicando que los resultados son significativos aún después de la corrección (figura 30dcha).

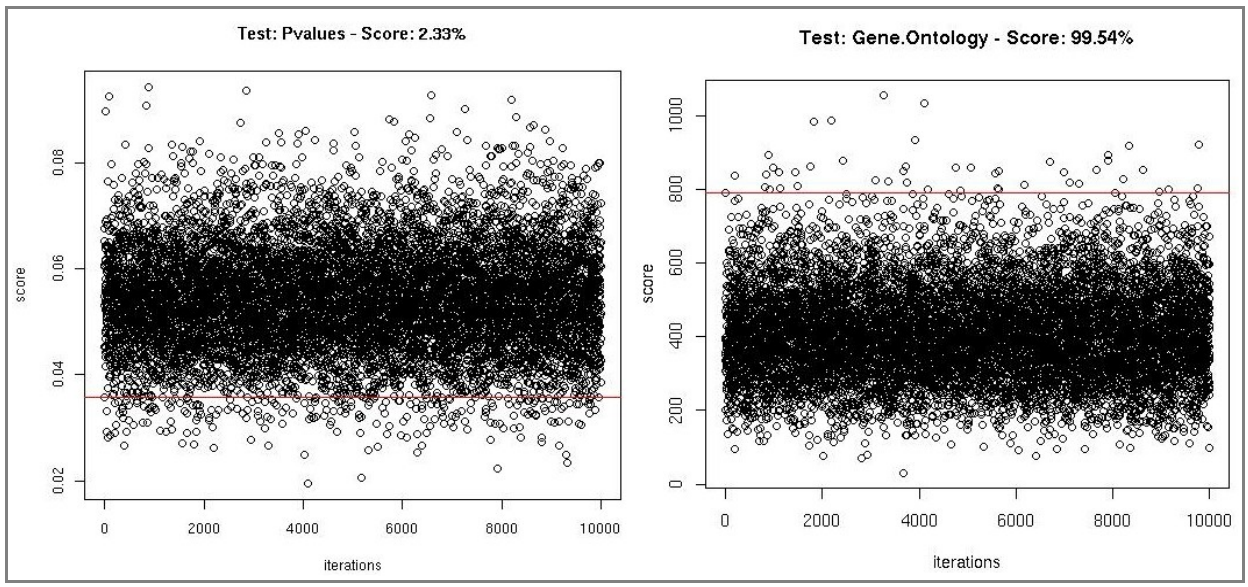


Figura 30. *Resultados obtenidos cuando se utilizan los muestreos obtenidos a partir del programa STRUCTURE.* ***izda:*** *distribución de scores obtenidos por permutación en el test de PV. Un 2.33% de las veces los scores obtenidos permutando las etiquetas de clases son menores que el score obtenido sin permutar (línea roja).* ***dcha:*** *distribución de scores obtenidos por permutación en el test de GO. El score de GO obtenido a partir de los datos originales (línea roja) es el 99.54% de las veces mayor que los obtenidos al azar.*

Estos datos indican que el método parece dar resultados preliminares prometedores cuando se incluyen las anotaciones funcionales de Gene Ontology. El método modela las interacciones entre marcadores según su localización en genes funcionalmente relacionados. Después de testar la aproximación en un conjunto de datos de pacientes con asma y de corregir la posible estratificación, se ha observado una correlación significativa (Pval = 0.0046) entre la enfermedad y el *score* funcional de GO.

4.2. Interpretación de los resultados del test GO

Además de poder dar una interpretación estadística, la introducción de las anotaciones de Gene Ontology en el test nos puede facilitar la interpretación biológica de los resultados. Así, podemos analizar los genes y términos GO obtenidos a partir los SNPs seleccionados para comprobar si se identifican genes o funciones relacionadas con la enfermedad de estudio y así valorar si la asociación encontrada podría ser causal y verificar si los resultados son relevantes.

Con el test de GO obtuvimos un conjunto de 34 genes asociados con los 100 SNPs de mejores estadísticos. Estos 34 genes forman el conjunto de genes asociados a la enfermedad en estudio. Para buscar las funciones en las que están implicados se obtuvieron los términos GO comunes entre todos los pares de genes que forman dicho conjunto.

De todos los posibles pares que forman ese conjunto de genes, hay 230 pares que comparten al menos un término GO, y por tanto esos son los pares que contribuyen a aumentar el *score S**. De todos ellos, los pares de genes que contribuyen más a ese aumento serán aquellos que tengan genes de estadísticos altos y compartan funciones comunes más especificas. En la tabla 6 se listan los pares cuya suma ponderada de estadísticos $(\chi^2_1 + \chi^2_2)$ GO_{niv} es más alta y por tanto son los pares que contribuyen más a aumentar el *S** funcional final.

χ^2 GEN 1	χ^2 GEN 2	Score	GO, nivel
17.461950	17.872760	8.314049	GO:0006468, niv.8
15.380947	17.872760	7.824402	GO:0006468, niv.8
15.380947	17.461950	7.727740	GO:0006468, niv.8
14.047195	17.872760	7.510578	GO:0006468, niv.8
14.047195	17.461950	7.413916	GO:0006468, niv.8
12.637103	17.872760	7.178791	GO:0006468, niv.8
12.637103	17.461950	7.082130	GO:0006468, niv.8
14.047195	15.380947	6.924269	GO:0006468, niv.8
11.444028	17.872760	6.898068	GO:0006468, niv.8
11.444028	17.461950	6.801407	GO:0006468, niv.8
12.637103	15.380947	6.592482	GO:0006468, niv.8
11.444028	15.380947	6.311759	GO:0006468, niv.8
12.637103	14.047195	6.278658	GO:0006468, niv.8
12.954845	13.378923	6.196181	GO:0006355, niv.8
12.637103	13.378923	6.121418	GO:0006355, niv.8
12.637103	12.954845	6.021635	GO:0006355, niv.8
11.444028	14.047195	5.997935	GO:0006468, niv.8
11.799879	13.378923	5.924424	GO:0006355, niv.8
11.799879	12.954845	5.824641	GO:0006355, niv.8
15.380947	17.405550	5.785852	GO:0005887, niv.6
15.340303	17.405550	5.778680	GO:0005887, niv.6
11.799879	12.637103	5.749878	GO:0006355, niv.8
11.444028	12.637103	5.666149	GO:0006468, niv.8
14.047195	17.405550	5.550484	GO:0005887, niv.6
11.481446	11.737892	5.463374	GO:0007207, niv.8
15.340303	15.380947	5.421397	GO:0005887, niv.6
12.375301	17.405550	5.255444	GO:0005887, niv.6
11.892701	17.872760	5.252728	GO:0006633, niv.6
14.047195	15.340303	5.186029	GO:0005887, niv.6
11.737892	17.405550	5.142960	GO:0005887, niv.6
11.490542	17.405550	5.099310	GO:0005887, niv.6
11.481446	17.405550	5.097705	GO:0005887, niv.6

Tabla 6. *Del conjunto de 34 genes se generan todos los pares de genes posibles. Para cada par se busca el termino GO común más específico, es decir, el termino GO común con nivel más alto. En la tabla se muestran los pares cuya suma ponderada de estadísticos (columna 3) es mayor. En la columna 1 se muestra el valor de χ^2 de uno de los genes del par, y en la columna 2 el χ^2 del segundo gen del par. En la columna 4 está el término GO compartido por los dos genes del par de mayor nivel.*

Estos pares están formados por 16 genes, que interesantemente no son los 16 genes con estadísticos más altos. Así por ejemplo el gen de estadístico $\chi^2 = 11.444028$ es el gen con menor estadístico de los 34 y sin embargo es uno de los genes que contribuye más a aumentar el *score* ya que comparte funciones muy específicas con otros genes asociados a SNPs de estadísticos altos.

De hecho al observar los estadísticos de esos 16 genes se encuentra que la media es 12.86, que es ligeramente menor que la media de los 34 genes (12.93). Esto indica que esos 16 genes están

contribuyendo a aumentar el *score* por el hecho de compartir GOs de niveles específicos, más que por el estadístico de los SNPs a los que están asociados.

Si observamos los términos GO compartidos, podemos ver que hay 5 términos que aparecen consistentemente entre esos pares (tabla 7).

Término GO	*GO ID, nivel*	*Descripción GO*
protein amino acid phosphorylation	GO:0006468 niv.8	*"The process of introducing a phosphoric group on to a protein".*
muscarinic acetylcholine receptor, phospholipase C activating pathway	GO:0007207 niv.8	*"The series of molecular signals generated as a consequence of a muscarinic acetylcholine receptor binding to its physiological ligand followed by the activation of phospholipase C and the subsequent release of inositol trisphosphate".*
integral to plasma membrane	GO:0005887 niv.6	*"Penetrating at least one phospholipid bilayer of a plasma membrane. May also refer to the state of being buried in the bilayer with no exposure outside the bilayer".*
regulation of transcription, DNA-dependent	GO:0006355 niv.8	*"Any process that modulates the frequency, rate or extent of DNA-dependent transcription".*
fatty acid biosynthesis	GO:0006633 niv.6	*"The formation from simpler components of a fatty acid any of the aliphatic monocarboxylic acids that can be liberated by hydrolysis from naturally occurring fats and oils. Fatty acids are predominantly straight-chain acids of 4 to 24 carbon atoms which may be saturated or unsaturated; branched fatty acids and hydroxy fatty acids also occur and very long chain acids of over 30 carbons are found in waxes".*

Tabla 7*. Términos GO que aparecen con mayor frecuencia en el conjunto de 34 genes. Para cada término GO aparece su nombre (columna1), su identificador y el nivel al que aparece anotado (columna 2) y su descripción (columna 3). Fuente http://www.geneontology.com*

En esta tabla se puede comprobar que entre los procesos compartidos en los genes que contribuyen más a aumentar el *score* funcional, hay un término GO, "***muscarinic acetylcholine receptor, phospholipase C activating pathway***", ampliamente relacionado con la enfermedad en estudio en la literatura (Fenech *et al*, 2001; Gosens *et al.*, 2006), ya que recientes investigaciones indican que la acetilcolina, actuando en los receptores muscarínicos pueden contribuir a la fisiopatología y patogénesis del asma (Gosens *et al.*, 2006).

El término GO "***protein amino acid phosphorylation***" aparece en los 10 pares de genes con mayor *score*. Si observamos en el árbol de términos GOs dónde se encuentra este termino, vemos que dos de sus hijos son "***JUN phosphorylation***" y "***peptidyl-arginine phosphorylation***", ambos ligados a la enfermedad en la literatura (Sousa *et al.*, 1999; Le Bellego *et al.*, 2006; Zimmermann *et al.*, 2003)

También el término "***integral to plasma membrane***" aparece frecuentemente en el conjunto de pares de genes de mayor *score*, y otra vez podemos encontrar que alguno de sus hijos, "***interleukin-3 receptor complex***" y "***interleukin-4 receptor complex***" están relacionados fuertemente con la enfermedad en la literatura (Battle *et al.*, 2007; López *et al.*, 2007).

Aunque estos dos últimos términos GO no están ligados directamente con la enfermedad de estudio, es interesante el hecho de que están directamente ligados, y por tanto relacionados, con otros procesos y componentes más específicos que si aparecen asociados a la enfermedad.

Se quiso comprobar si estos resultados aparentemente significativos en términos de funcionalidad fueron obtenidos por azar, en el sentido de que esos términos pudieran aparecer debido a un enriquecimiento de esos términos en el conjunto de datos original, no sólo en los genes asociados a los mejores 100 SNPs. Para ello se obtuvieron todos los genes asociados a todos los SNPs originales.

Se utilizó la herramienta FatiGO (Al-Shahrour *et al.*, 2004), la cual testea la distribución de términos GO entre dos grupos de genes por medio de un test de Fisher y ajusta los p-valores por FDR (*False Discovery Rate*). Con esta herramienta se buscaron los términos GO enriquecidos en el conjunto de 34 genes en comparación con el conjunto de todos los genes.

Al ejecutar la herramienta se encontró un enriquecimiento de los términos GO "***positive regulation of smooth muscle contraction***" y "***acetylcholine receptor signaling***". Este enriquecimiento refleja una mayor predominio relativo de esos términos entre los genes asociados a los SNPs de mejor estadístico (el 5.565% de esos genes tienen esos términos GO) que entre el conjunto completo de genes (ninguno tiene asociado esos términos, al menos en la versión de la base de datos utilizada). Estos dos términos están estrechamente ligados al término "***muscarinic acetylcholine receptor***" que se encontró con el test de GO, y a la enfermedad estudiada (Fenech *et al.*, 2001; Gosens *et al.*, 2006).

A este nivel de ontología también podemos ver que el término "***protein amino acid phosphorylation***", que se encontró con el método del test de GO también esta sobre-representado en el conjunto de genes ya que aparece en un 27.78% frente al 8.06% del total de genes (figura 31).

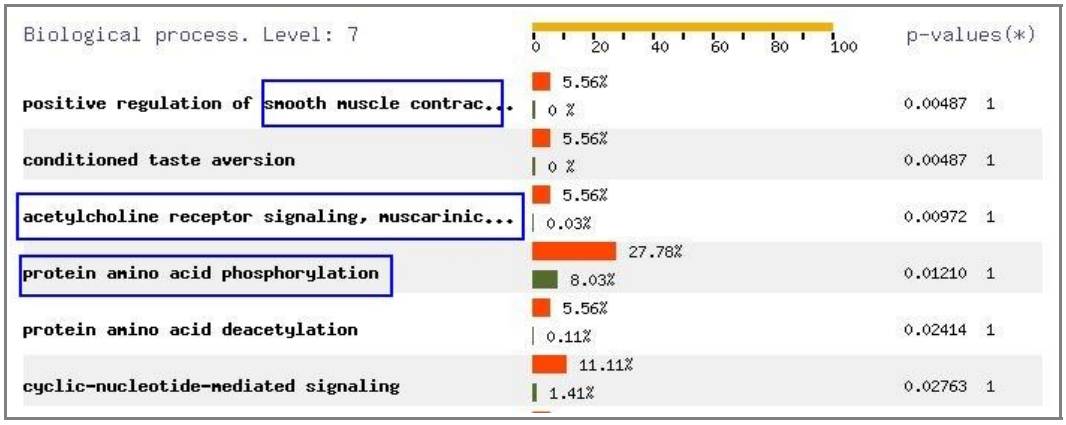


Figura 31. *Comparación de la distribución de términos GO entre el grupo de 34 genes (rojo) y el resto de genes (verde). A pesar de que el Pvalor corregido no es significativo, se puede observar que los términos "positive regulation of smooth muscle contraction" , "acetylcholine receptor signaling" y "protein amino acid phosphorylation" aparecen sobre-representados en el conjunto de genes asociado a los mejores 100 SNPs. Estos términos GO aparecen relacionados en la literatura con la enfermedad estudiada.*

Otro término que también se encontró con el test GO y que también aparece sobre-representado en el conjunto de 34 genes es "***integral to plasma membrane***" (41.18% vs. 15.58%) (figura 32).

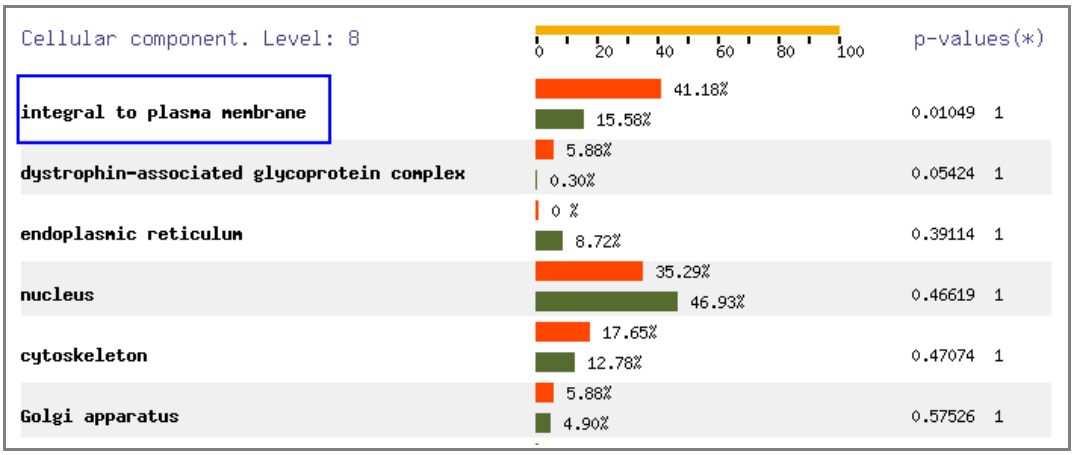


Figura 32. *El término "integral to plasma membrane" aparece sobre-representado en el conjunto de 34 genes asociados a la enfermedad. Este término aparece en el árbol jerárquico de GO conectado a varios complejos receptores de citoquinas relacionadas con la enfermedad en estudio.*

Si bajamos en el árbol jerárquico de GOs, además de los términos mencionados arriba, podemos encontrar otros términos sobre-representados en nuestro conjunto de genes asociados a los mejores 100 SNPs que están asociados a la enfermedad de estudio, como "***G-protein coupled receptor protein signaling pathway***" (figura 33), ya que es sabido que las variaciones genéticas en determinados

receptores acoplados a proteínas G están asociadas con un amplio espectro de predisposición a enfermedades respiratorias, incluyendo asma (Thompson *et al.*, 2006).

Figura 33. *Los términos "muscarinic acetylcholine receptor, phospholipase C activating pathway" y "G-protein coupled receptor protein signaling pathway" aparecen sobre-representados en el conjunto de 34 genes asociados a la enfermedad cuando se comparan con el total de genes.*

Por tanto, al analizar los términos GO o funciones que aparecen en los SNPs seleccionados y que como bloque presentan asociación, hemos encontrado términos relacionados con la enfermedad en estudio, con lo que este método podría utilizarse para confirmar el papel biológico de las asociaciones positivas.

Tests de interacción proteína-proteína y conservación

Por otra parte, los test de interacción proteína-proteína y de conservación no mostraron ningún resultado significativo, es decir, el conjunto de SNPs que mejor distingue entre casos y controles no parece ser un grupo de SNPs que se acumulen en zonas conservadas ni parece que estén implicados en interacciones físicas, por ejemplo formando complejos proteicos. Sin embargo, hay que mencionar que sólo se disponen de 38,223 entradas en la base de datos de interacciones, y muchos de los genes

asociados a esos SNPs no tenían ninguna anotación, por lo tanto no se puede descartar que no existan interacciones sino que no se han podido detectar y probablemente a medida que se vayan añadiendo anotaciones a las bases de datos, este test podrá proporcionar mayor información.

5

DISCUSIÓN

Los polimorfismos de un solo nucleótido o SNPs son sin duda un recurso muy valioso para investigar las bases genéticas de las enfermedades. Estos polimorfismos se han utilizado típicamente como marcadores en experimentos de mapeo genético y en estudios de asociación.

Con la introducción de las técnicas de genotipado a gran escala, el cuello de botella en este tipo de experimentos ha pasado a ser el manejo y análisis de la cantidad enorme de datos generados. La selección del conjunto óptimo de SNPs para los experimentos de genotipado no es una tarea fácil. Los SNPs óptimos tienen que ser los mejores marcadores posibles para enfermedades o rasgos que muchas veces son multigénicos, y hay que tener en cuenta factores como el **desequilibrio de ligamiento** (LD) y la **frecuencia alélica** del SNP (MAF) en la población de interés. Recientemente, la predicción del posible **efecto funcional** de los SNPs está teniendo cada vez más importancia como criterio de selección ya que constituye un factor importante para aumentar la sensibilidad de los estudios de asociación. De hecho, varias enfermedades complejas como la enfermedad de Alzheimer (Strittmatter y Roses, 1996) y la enfermedad de Crohn (Hugot *et al.*, 2001) se han asociado con SNPs funcionales, dando crédito a las estrategias que dan prioridad a marcadores candidatos basándose en una función predicha.

Recursos bioinformáticos para el análisis de polimorfismos funcionales

El número de polimorfismos funcionales en el genoma humano, que pueden predisponer a un individuo a enfermedades como diabetes, hipertensión, cáncer o afectar la progresión de la enfermedad, es desconocido, pero pueden ser decenas de miles. Identificar cuáles son funcionales entre los millones de polimorfismos que aparecen descritos hoy en día en bases de datos como dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP) es importante para la salud y para la investigación, pero claramente cada variación no se puede evaluar experimentalmente, sino que se necesitan herramientas bioinformáticas que valoren la probabilidad de funcionalidad basándose en los datos experimentales.

Hoy en día existen numerosas bases de datos y herramientas bioinformáticas con mucha información sobre SNPs. El catálogo principal de SNPs es la base de datos dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP), pero otras bases de datos más generales como Ensembl también incluyen en sus servidores esta información, que es procesada para ofrecer al usuario información más completa. Por ejemplo, el conocer el contexto genómico en el que se encuentra el SNP sirve como primer paso para delimitar los potenciales efectos funcionales de ese SNP (figura 34).

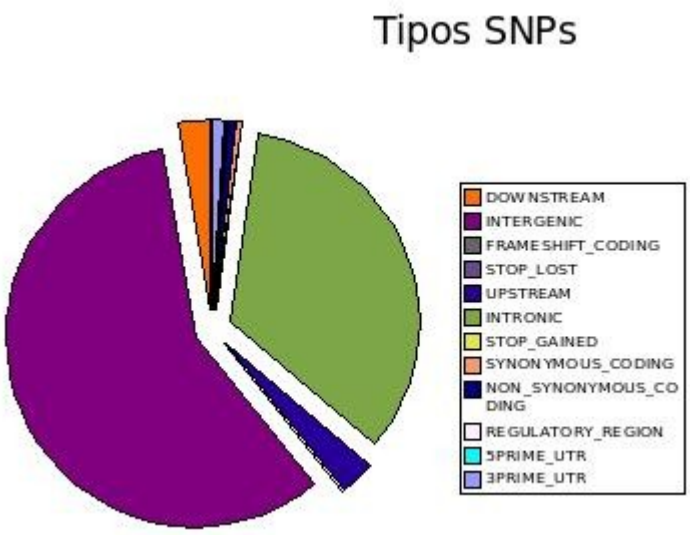


Figura 34. *Las bases de datos genómicas proporcionan información útil para un primer análisis de funcionalidad. En la figura se muestra la distribución de SNPs, contenida en Ensembl, según su tipo o localización en el genoma. La mayor parte de SNPs están en posiciones intergénicas e intrónicas, pero un pequeño porcentaje están localizados en las zonas río arriba y/o río abajo de genes y en zonas reguladoras. SNPs en esas regiones pueden seleccionarse para un análisis más detallado mediante programas bioinformáticos adecuados.*

Además de la base de datos de dbSNP, existen diferentes recursos disponibles en web, que recogen información de fenotipos asociados con SNPs, como *i)* 'The Human Gene Mutation Database' (HGMD), de la Universidad de Wales en Cardiff (Stenson et al., 2003), que clasifica a los SNPs de acuerdo con la lesión que causan (variantes de *splicing*, mutaciones terminadoras, pequeñas deleciones,...), y que en agosto de 2007 contiene descritas 53,181 mutaciones en 2,056 genes (http://www.hgmd.cf.ac.uk/ac/index.php), o *ii)* HGVbase, donde se detallan cómo, los SNPs y otros tipos de variación incluidos en la base de datos, están funcionalmente y físicamente relacionados con el gen más cercano, para facilitar los análisis de asociación genotipo-fenotipo.

Estas bases de datos son principalmente catálogos que recogen información sobre SNPs, más que herramientas para su selección. Afortunadamente existen hoy en día numerosos recursos disponibles para llevar a cabo análisis de las posibles consecuencias funcionales de los SNPs, como LS-SNP (http://alto.compbio.ucsf.edu/LS-SNP), PolyPhen (http://genetics.bwh.harvard.edu/pph/index.html) o SNPeffect (http://snpeffect.vib.be/index.php).

Sin embargo, las aproximaciones más comunes para la búsqueda de SNPs funcionales, incluidas las anteriores, se centran en el posible efecto de polimorfismos que causan cambio de aminoácido, a pesar de que las alteraciones de la regulación, nivel de expresión o *splicing* de genes pueden tener un importancia obvia en el fenotipo.

Obviamente la alteración funcional de codones muy conservados y sitios de *splicing*, que provocan una alteración de la estructura y función de la proteína, se detectan de una forma más fácil que las alteraciones de regiones reguladoras menos conservadas como promotores, potenciadores, silenciadores o intrones. Estas alteraciones son más complicadas de predecir, y sólo se conoce el mecanismo de acción de algunos de estos SNPs funcionales. Los SNPs reguladores pueden manifestarse de varias formas, incluyendo

- la alteración de la función de un elemento importante para la regulación normal
- una diferencia de afinidad de unión de una proteína
- o mediante la alteración de la función de un elemento que normalmente no participa en la regulación normal

En la tabla 9 se muestran algunos ejemplos de polimorfismos reguladores descritos en la literatura.

Localización	*Gen / enfermedad*	*Mecanismo*
TFBS	TNF / malaria	SNP en -376A introduce un sitio de unión a OCT1 que altera la expresión de TNF, asociado a un riesgo 4 veces mayor de susceptibilidad a malaria (Knight, 2005)
Promotor	α-globina / talasemia	Un SNP en la región promotora del gen de la α-globina (crom.16, 149709) crea un elemento similar a un nuevo promotor que interrumpe la expresión normal de los genes de la α-globina situados río abajo (De Gobbi *et al.*, 2006)
Sitios de *splicing*	ATP7A / síndrome de Menkes	Mutación en el sitio donador del exón 6 de ATP7A causa síndrome de Menkes, un síndrome letal relacionado con el metabolismo del cobre (Møller *et al.*, 2000)
ESE	MLH1 / cáncer colorectal hereditario no polipósico	Mutación en un ESE del exón 3 del gen MLH1 produce cáncer colorectal hereditario no-polipósico (McVety, 2006)
ISE	α-galactosidasa A / enfermedad de Fabry	Una mutación G->A en el intrón 4 del gen produce un aumento de *splicing* alternativo en el gen, produciendo una deficiencia en el catabolismo de glicoesfingolípidos (Ishii *et al.*, 2002)
ISI	TAU / demencia con Parkinsonismo	Una mutación en un ISE del intrón 11 causa la enfermedad por la alteración del *splicing* del exón 10 (D'souza y Schellenberg, 2000)
ESS	CD45 / esclerosis múltiple	Un polimorfismo en un ESS promueve el *splicing* del exón 4 de CD45 y su presencia se correlaciona con la susceptibilidad a desarrollar esclerosis múltiple (Lynch y Weiss, 2001)

***Tabla 9**. Ejemplos de polimorfismos funcionales en distintos elementos reguladores.*

Integración de recursos

A pesar de la gran cantidad de bases de datos y recursos bioinformáticos que existen para el

análisis funcional de regulación y *splicing* de genes (tabla 10), no hay muchas herramientas bioinformáticas que combinen toda la información disponible y proporcionen una herramienta integrada para el análisis funcional de SNPs reguladores.

Por esta razón, con la idea de desarrollar una aplicación que no sólo incluyese bases de datos y métodos de análisis de SNPs codificantes, sino que también integrase las diferentes posibilidades disponibles para el análisis de SNPs reguladores, durante esta tesis se ha desarrollado un conjunto de herramientas (Conde *et al.*, 2004; Conde *et al.*, 2005; Conde *et al.*, 2006), que se han resumido finalmente en una sola, PupaSuite, que permite la selección de conjuntos óptimos de SNPs orientado a estudios de asociación a gran escala. La herramienta incorpora no sólo métodos de análisis de funcionalidad sino también métodos para el cálculo de parámetros de LD, bloques de haplotipos y tags, así como información sobre MAF en distintas poblaciones. Además, la posibilidad de aplicar los distintos filtros y visualizar los datos en un formato gráfico la hacen una herramienta intuitiva y fácil de usar.

Herramientas	*Descripción*	*Referencia*
Ensembl	Base de datos genómica	Hubbard *et al.*, 2007
UCSC	Base de datos genómica	Kuhn *et al.*, 2007
PromoterInspector	Predicción de promotores	Scherf *et al.*, 2000
Eponine	Predicción de sitios de inicio de transcripción	Down y Hubbard, 2002
CpGPlot	Predicción de islas CpG	Larsen *et al.*, 1992
EPD	Base de datos de promotores eucariotas	Schmid *et al.*, 2006
TRANSFAC®	Base de datos de factores de transcripción	Wingender *et al.*, 2000
JASPAR	Base de datos de factores de transcripción	Sandelin *et al.*, 2004
Match™	Predicción de TFBSs	Kel *et al.*, 2003
ESEfinder	Predicción de ESEs	Cartegni *et al.*, 2003
RESCUE-ESE	Predicción de ESEs	Fairbrother *et al.*, 2002
GeneId	Predicción de genes	Guigó, 1998
GENSCAN	Predicción de genes	Burge y Karlin, 1997
miRBase	Base de datos de microRNAs	Griffiths-Jones *et al.*, 2006

***Tabla 10**. Ejemplos de herramientas para el análisis funcional de elementos de regulación génica y splicing. En la tabla aparecen recursos genómicos más generales, como los buscadores de Ensembl y UCSC, y herramientas mas específicas para la identificación de promotores, sitios de unión a factores de transcripción, potenciadores de splicing exónicos, microRNAs y herramientas de predicción de genes que permiten la detección de sitios de splicing.*

En los últimos años se han ido desarrollando otras herramientas similares, como por ejemplo PromoLig (Zhao *et al.*, 2004), SNPselector (Xu *et al.*, 2005) y FastSNP (Yuan *et al.*, 2006). A pesar de que éstas proporcionan una plataforma útil para el análisis de SNPs, el número de análisis que realizan es menor, por lo que PupaSuite es una de las herramientas más completas para el análisis y selección de SNPs. La utilidad de esta herramienta se demuestra por el hecho de formar parte de la plataforma de genotipado del Centro Nacional de Genotipado (CeGen, http://www.cegen.org), como herramienta de soporte en la selección de SNPs de regiones o genes de interés previamente fijados por el investigador.

Análisis bioinformático de SNPs

Mediante el uso los distintos métodos de predicción incorporados en PupaSuite se ha realizado un estudio de todos los SNPs descritos en el genoma humano, y se han encontrado un total de 499,640 SNPs en distintos elementos reguladores que podrían tener efectos en mecanismos importantes como transcripción y *splicing*. Esto supone que un 5% del total de SNPs podrían ser SNPs reguladores.

Entre todos estos SNPs con potencial efecto regulador se encuentran:

- 1,122 SNPs en sitios de *splicing*
- 17,957 SNPs en silenciadores de *splicing* exónicos
- 91,613 SNPs en potenciadores de *splicing* exónicos
- 299,947 SNPs en secuencias capaces de formar triples
- 95,255 SNPs situados en sitios de unión a factores de transcripción

Probablemente estas cifras estén hinchadas por un número grande de falsos positivos. Para disminuir el número de falsos positivos se puede adoptar un criterio de conservación, ya que en general está aceptado que mutaciones en esas regiones se eliminan por selección natural y por tanto no son fenotípicamente neutrales (Asthana *et al.*, 2007).

En cuanto a los SNPs codificantes, se ha encontrado que aproximadamente un 13% de los nsSNPs tienen un potencial efecto patológico teniendo en cuenta la estima de la presión selectiva a nivel de codón. Si añadimos las predicciones obtenidas a través del programa Pmut (Ferrer-Costa *et al.*, 2002, 2004, 2005) y de la base de datos SNPeffect (Reumers *et al.*, 2005, 2006), que también se incluyen en PupaSuite, encontramos que el porcentaje total de nsSNPs predichos como patológicos, por cualquiera de los 3 métodos, es de un 28% (17,242 nsSNPs, figura 35).

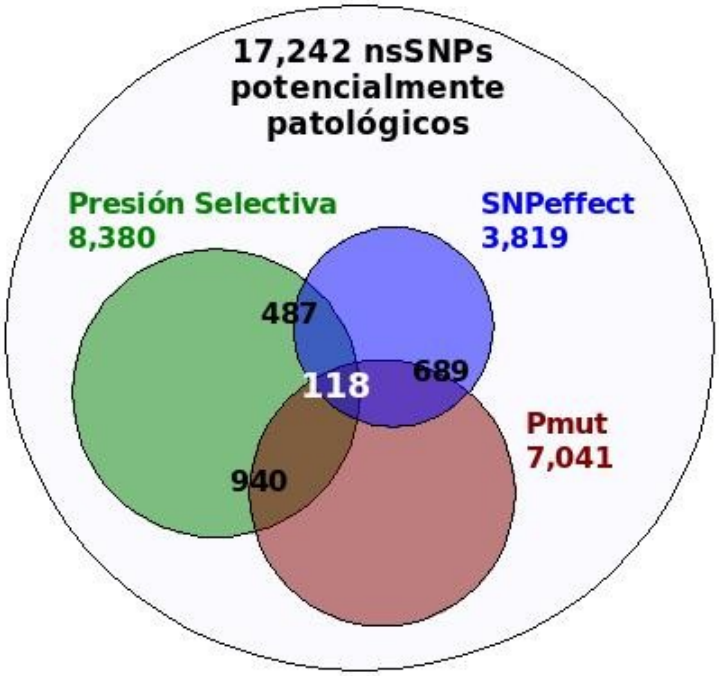


Figura 35. *Aproximadamente un 28% (17,242) de los SNPs codificantes no sinónimos tienen potencial efecto patológico teniendo en cuenta la estima de la presión selectiva a nivel de codón (8,380 SNPs se predicen como patológicos por este método) y las predicciones de Pmut (predice 3,819 SNPs patológicos con un nivel de fiabilidad mayor de 4, en una escala de 0 a 9) y SNPeffect (7,041 SNPs producen cambios significativos en el procesamiento celular y estructura proteica). Existen 118 nsSNPs que se predicen como patológicos por los 3 métodos de predicción de funcionalidad de nsSNPs incluidos en PupaSuite.*

Este porcentaje total se asemeja al porcentaje, 25%-30%, de nsSNPs patológicos predicho por la mayoría de métodos (Ng y Henikoff, 2006). Sin embargo, de los 17,242 SNPs, solamente 118 de ellos (un 0.7%) son predichos como patológicos por los 3 métodos de predicción incluidos en PupaSuite. Este número tan bajo se debe a la dificultad de encontrar predicciones de funcionalidad consenso que incluyan múltiples métodos. Ya que son métodos distintos que parten de distinta información, la cobertura de los métodos es distinta. Así por ejemplo la información que detalla residuos implicados en interacciones proteínas-proteínas y unión de ligandos es todavía escasa y por tanto no todos los SNPs podrán analizarse con métodos que necesiten esta información, como parte de los métodos incluidos en SNPeffect. Por otra parte SNPs situados en regiones no cubiertas por alineamientos con secuencias ortólogas no podrán analizarse por el método de presión selectiva. Recientemente Burke y colaboradores (Burke *et al.*, 2007) realizaron un estudio en el que compararon y contrastaron distintos métodos de predicción sobre 21,471 nsSNPs provenientes de dbSNP. Encontraron que un 10% de los SNPs son predichos como patológicos por LS-SNP (Karchin *et al.*, 2005), y que un 5% lo son por PolyPhen (Ramensky *et al.*, 2002), pero solamente un 1% son predichos como dañinos por ambos métodos, y al añadir las predicciones de SIFT (Ng y Henikoff, 2003) ese porcentaje baja al 0.6%.

A medida que se hagan más completas las anotaciones de las bases de datos, se podrá mejorar la cobertura de los distintos métodos de predicción y así se podrá mejorar la interpretación de los datos obtenidos. Mientras tanto, la predicción de los efectos causados por SNPs es una cuestión importante que está aun sin resolver, y la combinación de los resultados obtenidos de la valoración de distintas propiedades (propiedades de secuencia, estructurales, sitios funcionales...) a partir de distintos métodos es algo muy atractivo que hace que herramientas como PupaSuite sean recursos muy valiosos para la comunidad científica.

Hipótesis de CD/CV o variantes raras

Un debate existente en genética molecular es el de conocer cuál es la contribución de las variantes raras a aumentar del riesgo a desarrollar una enfermedad común. Existen dos teorías contrapuestas, una de ellas apoya que las variantes comunes, con alelos relativamente frecuentes, son las que más contribuyen a predisponer a los individuos a la enfermedad y/o influencian sus respuestas a fármacos. Es la teoría de la enfermedad común/variante común (CD/CV) (Reich y Lander, 2001). Sin embargo en algunos estudios se observa que existe una asociación inversa entre las MAF de nsSNPs y la predicción funcional dañina de esos SNPs (Rudd *et al.*, 2005), es decir, los alelos que son funcionalmente dañinos tienden a ser seleccionados en contra y no existirán con una elevada frecuencia.

Para ver la relación entre las MAF y las predicciones de funcionalidad encontradas anteriormente, se buscaron los datos de frecuencias poblacionales de todos los SNPs predichos como funcionales y se compararon con el resto de SNPs de las mismas regiones que no se catalogaron como funcionales. Se observó que en general los SNPs predichos como funcionales aparecen con una frecuencia menor y son más específicos de población, algo que concuerda con la teoría de la selección purificadora de reducir la frecuencia de alelos dañinos. Sin embargo, con la excepción de los SNPs situados en sitios de *splicing*, éstas frecuencias alélicas son siempre mayores del 1%, es decir, no llegan a ser variantes raras sino que son variantes comunes, algo de esperar ya que esas frecuencias se obtuvieron de HapMap, donde la gran mayoría de las variaciones anotadas son comunes. Por tanto de estos resultados no se pueden extraer conclusiones que permitan decantarse por una hipótesis o por otra, y tal vez los más probable sea que los SNPs comunes catalogados como funcionales aumenten el riesgo a desarrollar una enfermedad pero no sean por si solos suficiente para causarla, y se necesiten otros factores multigénicos y medioambientales para causar la enfermedad. Es creíble pensar que las enfermedades comunes estén controladas por mecanismos genéticos más complejos caracterizados

por la acción conjunta de varios genes, cada uno con un pequeño efecto marginal, tal vez porque la selección natural haya eliminado aquellos con efectos muy grandes.

Reflexión sobre las predicciones

¿Cuantos de los SNPs predichos como funcionales están realmente actuando como tal? La mayoría de los elementos reguladores conocidos se han detectado empleando metodologías *in vitro*, con lo que las conclusiones derivadas de este tipo de predicciones bioinformáticas deben tomarse con precaución. Debido a la enorme complejidad de genes, tránscritos y proteínas, existen infinitas posibilidades de formular hipótesis sobre la funcionalidad de los SNPs y probablemente es posible asignar un efecto potencialmente dañino casi a todos los SNPs. Pero claramente el genoma humano no contiene millones de mutaciones potencialmente dañinas, por tanto es importante tratar las predicciones *in silico* con prudencia, ya que las predicciones son precisamente eso, predicciones, y para caracterizar el mecanismo molecular de un SNP con potencial efecto funcional puede ser necesario una combinación de métodos bioinformáticos con un posterior seguimiento en el laboratorio.

Futuras mejoras

En el momento de escribir esta tesis, ya se están empezando a incluir futuras mejoras para la herramienta.

- Se está desarrollando una aproximación complementaria para la identificación de los sitios de unión a factores de transcripción, utilizando las matrices de pesos de la base de datos JASPAR (Sandelin *et al.*, 2004), cuyos modelos derivan de 81 PWMs verificadas biológicamente y los programas MatScan y Meta (http://genome.imim.es). MatScan es una programa de búsqueda de sitios de unión en secuencias genómicas. Debido a que el programa MatScan no permite un corte para minimizar los falsos positivos, se pretende utilizar el programa Meta para filtrar los resultados mediante la búsqueda de coincidencias de TFBSs en genes ortólogos en ratón.
- También se incluirán las predicciones de GeneID (Guigó, 1998). Éste es un programa de predicción de genes en secuencias genómicas anónimas diseñado siguiendo una estructura jerárquica en la que el primer paso es la utilización de PWMs para predecir y dar un *score* a los sitios de *splicing* a lo largo de la secuencia. Aunque en principio es un programa de predicción *in silico* de genes, se puede utilizar para la búsqueda de nuevos sitios de *splicing*.

- También se prevé la inclusión de un nuevo método de predicción de nsSNPs asociados a enfermedad por métodos evolutivos (Capriotti *et al.*, 2007).
- Además, elementos reguladores adicionales como las caja TATA, islas CpG, elementos repetitivos o microRNAs y sus dianas son elementos donde la presencia de SNPs puede tener potenciales consecuencias funcionales. Por eso, estos elementos se tendrán en cuenta en futuras versiones de la herramienta.
- Finalmente, ya se están empezando a realizar las predicciones en los genomas de ratón y rata para su inclusión en la siguiente versión de Pupasuite. De esta forma la herramienta puede ayudar a comprender mejor la diversidad funcional en los distintos genomas.

Variaciones de número de copia

Los SNPs son probablemente el tipo de variación genética más estudiada debido a su prevalencia en el genoma, pero no son la única variación genética existente en el genoma humano. Ya desde el comienzo de la citogenética se han podido observar bajo el microscopio variaciones en el número de cromosomas y reordenamientos que en muchos casos se han podido asociar a enfermedades, como la copia adicional del cromosoma 21 en el síndrome de Down. Por tanto, la variación genética en humanos varía entre un cambio de una base, hasta diferencias cromosómicas de varias megabases detectables por microscopio.

Recientemente nuestra visión de la variación genética se ha extendido por la observación de abundantes variaciones en el número de copia de segmentos de DNA submicroscópicos (CNVs). Incluso algunos estudios parecen indicar que este número de CNVs en el genoma humano podría superar al de SNPs (Lee, 2005). Debido a que las CNVs a menudo abarcan genes enteros, es obvio que pueden jugar papeles importantes tanto en enfermedades como en respuesta a drogas, y comprender los mecanismos de la formación de CNVs puede también ayudar a entender mejor la evolución del genoma humano. Varias instituciones han empezado a desarrollar bases de datos de CNVs asociadas con condiciones clínicas, como la 'Database of Genomic Variants' (http://projects.tcag.ca/variation), obtenida a partir del estudio de aproximadamente 1,000 genomas de individuos sin ningún fenotipo de enfermedad aparente, el 'Human Genome Structural Variation Project' (http://humanparalogy.gs.washington.edu/structuralvariation) y la 'Database of Chromosome Imbalances in Phenotypes Using Ensembl Resources, DECIPHER' (http://www.sanger.ac.uk/PostGenomics/decipher), creada por el Wellcome Trust Sanger Institute.

La nueva generación de tecnologías basadas en *microarrays* de DNA permitirá la detección de

nuevas CNVs a medida que se analicen muestras de poblaciones de todo el mundo, y puede que en menos de un año la cantidad de datos aumente en varios ordenes de magnitud (Scherer *et al.*, 2007), y por tanto, es imprescindible el desarrollo de herramientas que permitan el almacenamiento, procesado y análisis de los datos generados.

Además de la identificación precisa de la región que tiene un número de copia alterado, sería deseable poder analizar la relación de los CNVs con los cambios de expresión génica en esas zonas y entender cual es el efecto funcional, a nivel molecular, que puede ayudar a interpretar la enfermedad o fenotipo estudiado. Ésto, aunque es importante, es un aspecto que se pasa por alto en la mayoría de las herramientas para el análisis de CNVs.

En ese sentido, durante esta tesis se ha desarrollado una herramienta web llamada ISACGH (Conde *et al.*, 2007a, 2007b), que permite simultáneamente el estudio de CNVs mediante *arrays* de CGH, sus efectos en la expresión génica y el posible impacto funcional de esa alteración cromosómica. Su inclusión en el paquete GEPAS además facilita los procesos de normalización, transformación de datos y otros análisis como la expresión diferencial, *clustering*, etc.

Análisis de datos de genotipado

Desde la finalización de la secuenciación del genoma humano, los estudios de asociación a gran escala se han considerado como una gran promesa para estudiar las bases genéticas de las enfermedades humanas. El progreso que se está realizando en el genotipado de SNPs y la disponibilidad de recursos como HapMap están haciendo posible realizar estudios de asociación de genomas enteros, como lo demuestra el reciente estudio de asociación realizado por el consorcio WTCCC sobre 14,000 casos y 30,000 controles para el estudio de 7 enfermedades comunes (The Wellcome Trust Case control Consortium, 2007). Sin embargo, a pesar de los éxitos que producen los estudios a escala genómica, no es tan obvio cómo analizar los datos de forma productiva. Se necesitan miles de SNPs para testar de manera eficiente la variabilidad genética del genoma, lo que necesariamente supone una corrección del testeo múltiple y un número muy elevado de muestras para poder detectar las señales, relativamente débiles, esperadas en enfermedades complejas. Además, la presencia de estratificación en las muestras es uno de los factores más importantes que llevan a asociaciones y conclusiones erróneas. Todos estos problemas (errores de tipo I, tamaño de muestras, estratificación, errores de genotipado...) han llevado al desarrollo de cada vez más y mejores métodos bioinformáticos y estadísticos que permiten, al menos parcialmente, solventarlos.

Por otra parte, la mayor parte de los métodos analíticos consideran cada marcador genético de

forma individual, pero cada vez hay mas evidencia, gracias a organismos modelos y a estudios humanos, que sugieren que las interacciones entre loci contribuyen en gran medida a los rasgos complejos. La complejidad de muchas enfermedades puede surgir por el hecho de que muchos factores genéticos (y medioambientales) pueden interaccionar unos con otros de forma casi impredecible, de forma que la asociación entre el fenotipo de la enfermedad y cualquier factor tomado individualmente puede se imperceptible. Si las enfermedades complejas están influenciadas por las interacciones entre múltiples loci y por tanto si el riesgo asociado a un locus esta influenciado por el genotipo de otro locus, los análisis de marcadores individuales no detectarán esa asociación. Centrarse en genes individuales no proporciona una imagen global de todos los SNPs que pueden participar en el mismo proceso celular o patológico. Para entender como distintos factores heterogéneos pueden llevar a fenotipos patológicos similares es necesario tener anotaciones funcionales de niveles más altos. Estas anotaciones deberían incluir por ejemplo rutas metabólicas, ontologías, interacciones de proteínas, etc., que se combinen con la información genotípica en los métodos de análisis de estudios de asociación.

Integración de datos de genotipado con información funcional

Durante esta tesis se han explorado formas de utilizar los datos de genotipado en combinación con información funcional en el contexto de estudios de asociación. Para calcular si la diferencia en la variación genética entre casos y controles es estadísticamente significativa se utilizan los estadísticos individuales de cada SNP para ordenarlos según su probabilidad a estar asociados a la enfermedad. El conjunto de los mejores marcadores (en el sentido de que distinguen mejor entre casos y controles) se combina y se añade información biológica (interacciones proteína-proteína, grado de conservación entre especies o descripción de lo productos génicos a través de Gene Ontology) para obtener un score funcional cuya significación estadística se evalúa mediante permutaciones.

Uno de los métodos que parece dar resultados preliminares prometedores es la asociación de los términos de Gene Ontology con un panel de casos y controles de pacientes con asma. La aproximación utilizada es similar a los métodos *Two Steps* ya que en la primera parte se reduce el número de marcadores a un número más pequeño que contribuyen más a diferenciar entre los casos y controles. De ese subgrupo de marcadores se obtuvo una lista no redundante de 34 genes, teniendo en cuenta la correlación entre los marcadores a la hora de asociar genes a SNPs. En una segunda parte se moldearon las interacciones entre pares de genes mediante el uso de la información disponible en GO y se proporcionó una significación estadística mediante el test de permutación. Se observó una

correlación significativa (p=0.0046) entre el score de GO y la enfermedad, y el posterior análisis de los términos encontrados parecen confirmar el papel biológico de esa asociación, ya que los genes candidatos encontrados están enriquecidos en términos como "*muscarinic acetylcholine receptor*" o "*positive regulation of smooth muscle contraction*", términos asociados a la enfermedad de estudio en la literatura.

El valor de esta aproximación probablemente aumentará al aumentar y mejorar las anotaciones genómicas disponibles; a medida que el conocimiento existente sobre interacciones proteína-proteína, rutas y ontologías vaya aumentando, este tipo de metodologías podrá facilitar la identificación de interacciones de riesgo importantes. En el futuro se prevé aplicar esta estrategia en otros conjuntos de datos con el objetivo de replicar los resultados y explorar variaciones del método para investigar interacciones más complejas que impliquen más de dos genes.

6

CONCLUSIONES

De los resultados expuestos se pueden extraer las siguientes conclusiones:

1. La bioinformática, en el campo de la investigación biomédica, tiene dos retos principales, el análisis de nuevos datos obtenidos por tecnologías de alto rendimiento, y la combinación e integración de diferentes tipos de datos para conseguir una visión más completa de la enfermedad. En esta tesis se introducen distintas herramientas que pretenden facilitar la consecución de esos retos en el campo de los polimorfismos genéticos.
2. Se han utilizado distintos métodos bioinformáticos para la identificación de polimorfismos genéticos funcionales en humano. El análisis a escala genómica muestra que aproximadamente un 5% de los polimorfismos podrían tener implicaciones funcionales a nivel transcripcional y que un 28% de los SNPs codificantes no sinónimos podrían tener un efecto dañino en la función de la proteína.
3. Se observó que en general los SNPs predichos como funcionales aparecen con una frecuencia menor y son más específicos de población. Con la excepción de los SNPs situados en sitios de *splicing*, éstas frecuencias alélicas son siempre mayores del 1%, ya que esas frecuencias se obtuvieron de HapMap, donde la gran mayoría de las variaciones anotadas son comunes.
4. Se han explorado métodos de análisis de datos de genotipado en los que se combinan estos datos con información biológica (como interacciones proteína-proteína o anotaciones de Gene Ontology) en el contexto de estudios de asociación. La inclusión de esta información puede facilitar el análisis e interpretación de los datos de una forma más intuitiva.
5. El método en el que se incluyen las anotaciones de Gene Ontology mostró resultados prometedores cuando se aplicó a un estudio de casos/controles con individuos afectados de asma, encontrándose una asociación significativa entre el *score* y la enfermedad ($p=0.0046$).

Z

BIBLIOGRAFÍA

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. (2004). "Fatigo: a web tool for finding significant associations of Gene Ontology terms with groups of genes". *Bioinformatics*. **20**, 578-580.

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. (2001). "Genome-wide scans of complex human diseases: true linkage is hard to find". *Am. J. Hum. Genet*. **69**, 936-950.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Res.* **25**, 3389–3402.

Arbiza L, Dopazo J, Dopazo H. (2006). "Positive selection, relaxation and acceleration in the evolution of the human and chimp genomes". *PLoS Comp. Biol*. **2(4)**, e38.

Ardlie KG, Kruglyak L, Seielstad M. (2002). "Patterns of linkage disequilibrium in the human genome". *Nat. Rev. Genet*. **3(4)**, 299-309.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. (2000) "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". *Nature Genet*. **25**, 25–29.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. (2007). "Widely distributed noncoding purifying selection in the human genome". *Proc Natl Acad Sci USA*. **104(30)**, 12410-5.

de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. (2005). "Efficiency and power in genetic association studies". *Nature Genetics*. **37**, 1217–1223.

Balding DJ (2006). "A tutorial on statistical methods for population association studies". *Nat. Rev. Genet*. **7(10)**, 781-791.

Bao L, Zhou M, Cui Y. (2005). "nsSNPAnalyzer: identifying disease-associated non-synonymous single nucleotide polymorphisms". *Nucleic Acids Res*. **33**, W480-2.

Baralle D, Baralle M. (2005). "Splicing in action: assessing disease causing sequence changes". *J. Med. Genet*. **42(10)**, 737-748.

Barnes MR, Gray IC. (2003). "Bioinformatics for Geneticists". John Wiley & Sons Ltd., 422 pp.

Barrett JC, Fry B, Maller J, Daly MJ. "Haploview: analysis and visualization of LD and haplotype maps". *Bioinformatics*. **21(2)**, 263-5.

Battle NC, Choudhry S, Tsai HJ, Eng C, Kumar G, Beckman KB, Naqvi M, Meade K, Watson HG, Lenoir M, *et al*. (2007). "Ethnicity-specific gene-gene interaction between IL-13 and IL-4Ralpha among African Americans with asthma". *Am J Respir Crit Care Med*. **175(9)**, 881-7.

Beheshti B, Park PC, Braude I, Squire JA. (2002). "Microarray CGH". *Methods Mol Biol.* 204, 191-207.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. (2004). "Ultraconserved Elements in the Human Genome". *Science*. **304**, 1321-1325.

Benjamini Y, Hochberg Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *J. Roy. Statist. Soc. Ser. B*. **57**, 289-300.

Benjamini Y, Hochberg Y (2000). "The adaptive control of the false discovery rate in multiple hypotheses testing". *J. Behav. Educ. Statist*. **25**, 60-83.

Borecki IB, Suarez BK. (2001). "Linkage and association: basic concepts". *Adv. Genet*. **42**, 45–66.

Botstein D, White RL, Skolnick M, Davis RW. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms". *American Journal of Human Genetics*. **32(3)**, 314–331.

Botstein D, Risch N. (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease". *Nature Genet.* **33**, 228–237.

Brookes AJ. (1999). "The essence of SNPs". *Gene*. **234(2)**, 177-86.

Bucher P. (1990). "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences". *J. Mol. Biol*. **212**, 563– 578.

Buckland PR. (2006). "The importance and identification of regulatory polymorphisms and their mechanisms of action". *Biochim. Biophys. Acta.* **1762,** 17–28.

Buratti E, Baralle M, Baralle FE. (2001). "Defective splicing, disease and therapy: searching for

master checkpoints in exon definition". *Nucleic Acids Res*. **29(1)**, 308-11.

Bureau A, Dupuis J, Faslls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. (2005), "Identifying SNPs predictive of phenotype using random forests". *Genet Epidemiol*. **28(2)**, 171-82.

Burge C, Karlin S. (1997). "Prediction of complete gene structures in human genomic DNA". *J. Mol. Biol.* **268,** 78-94.

Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL. (2007). "Genome bioinformatic analysis of nonsynonymous SNPs". BMC Bioinformatics. **8(1)**, 301.

Burke TW, Kadonaga JT. (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters". *Genes & Dev*. **10**, 711–724.

Burset M, Seledtsov IA, Solovyev VV. (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes". *Nucleic Acids Res*. **28(21)**, 4364-75.

Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS. (2004). "Bayesian approach to discovering pathogenic SNPs in conserved protein domains". *Human Mutat*. **24(2)**, 178–184.

Campbell H, Rudan I. (2002). "Interpretation of genetic association studies in complex disease". *The Pharmacogenomics Journal.* **2**, 349-360.

Capon F, Allen MH, Amee M. (2006). "A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups". *Human Molecular Genetics*. **13(20)**. 2361–2368.

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Martí-Renom MA. (2007) "The use of estimated evolutionary strength at the codon level improves the prediction of disease related protein mutations in human". *Human Mutation. In Press*

Cardon LR, Bell JI. (2001). "Association study designs for complex diseases". *Nat. Rev. Genetics.* **2**, 91-9.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. (2004). "Mapping complex disease loci in whole-genome association studies". *Nature*. **429(6990)**, 446-52.

Carmel I, Tal S, Vig I, Ast G. (2004). "Comparative analysis detects dependencies among the 50

splice-site positions". *RNA*. **10**, 828–840.

Cartegni L, Chew SL, Krainer AR. (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing". *Nature Rev. Genet.* **3**, 285–298.

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. (2003). "ESEfinder: a web resource to identify exonic splicing enhancers". *Nucleic Acids Res*. **31**, 3568– 3571.

Chakravarti A. (1999). "Population Genetics – making sense out of sequence". *Nat. Genet*. **21**, 56-60.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, *et al*. (2005). "Population structure, differential bias and genomic control in a large-scale, case-control association study". *Nature Genetics*. **37**, 1243 - 1246.

Collins FS, Green ED, Guttmacher AE, Guyer MS; US National Human Genome Research Institute. (2003). "A vision for the future of genomics research". *Nature*. **422(6934)**, 835-47.

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. (2004). "PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level". *Nucleic Acids. Res*. **32**, W242-W248.

Conde L, Vaquerizas JM, Ferrer-Costa C, Orozco M., Dopazo J. (2005). "PupasView: a visual tool for selecting suitable SNPs, with putative pathologic effect in genes, for genotyping purposes". *Nucleic Acids Res*. **33**, W501-5.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J. (2006). "PupaSuite: finding functional SNPs for large-scale genotyping purposes". *Nucl Acids Res*. **34**, W621-W625.

Conde L, Montaner D, Burguet-Castell J, Tarraga J, Medina I, Al-Shahrour F, Dopazo J. (2007a). "ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling". *Nucleic Acids Res*. **35**, W81-5.

Conde L, Montaner D, Burguet-Castell J, Tarraga J, Al-Shahrour F, Dopazo J. (2007b). "Functional profiling and gene expression analysis of chromosomal copy number alterations". *Bioinformation*. **1(10)**, 432-435.

Croiseau P, Génin E, Cordell HJ. (2007). "Dealing with missing data in family-based association studies: a multiple imputation approach.". *Hum. Hered*. **63(3-4)**, 229-38.

Culverhouse R, Klein T, Shannon W. (2004). "Detecting epistatic interactions contributing to quantitative traits". *Genet Epidemiol*. **27**, 141–152.

Culverhouse R. (2007). "The use of the restricted partition method with case-control data". *Human Heredity*. **63(2)**, 93-100.

Dai JY, Ruczinski I, LeBlanc M, Kooperberg C. (2006). "Imputation methods to improve inference in SNP association studies". *Genet Epidemiol*. **30(8)**, 690-702.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, *et al*. (2002). "A first-generation linkage disequilibrium map of human chromosome 22". *Nature*. **418**, 544-8.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, *et al*. (2006). "A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter". *Science*. **312(5777)**, 1215-7.

Devlin B, Risch N. (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping". *Genomics*. **29(2)**, 311-22.

Devlin B, Roeder K, Wasserman L. (2001). "Genomic control, a new approach to genetic-based association studies". *Theor Popul Biol*. **60**, 155-66.

Dobson RJ, Munroe PB, Caulfield MJ, Saqi MA. (2006). "Predicting deleterious nsSNPs: an analysis of sequence and structural attributes". *BMC Bioinformatics*. **7**, 217.

Down TA, Hubbard TJ. (2002). "Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA". *Genome Res*. **12**, 458-461.

D'Souza I, Schellenberg GD. (2000). "Determinants of 4-repeat tau expression. Coordination between enhancing and inhibitory splicing sequences for exon 10 inclusion". *J Biol Chem*. **275(23)**, 17700-9.

Eisen M, Spellman P, Brown P, Botstein D. (1998). "Cluster analysis and display of genome-wide expression patterns". *PNAS*. **95**, 14863-14868.

Elston RC, Spence MA. (2006). "Advances in statistical human genetics over the last 25 years". *Stat. Med*. **25(18)**, 3049-80.

ENCODE Project Consortium (2004). "The ENCODE (ENCyclopedia Of DNA Elements) project". *Science*. **306**, 636-640.

Ewens WJ, Spielman RS. (2005). "What is the significance of a significant TDT?". *Human Heredity*. **60(4)**, 206–210.

Excoffier L, Heckel G. (2006). "Computer programs for population genetics data analysis: a survival guide". *Nature Reviews Genetics*. **7**, 745-758.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. (2002). "Predictive identification of exonic splicing enhancers in human genes". *Science*. **297(5583)**, 1007-13.

Fairbrother WG, Holste D, Burge CB, Sharp PA (2004). "Single Nucleotide Polymorphism–Based Validation of Exonic Splicing Enhancers". *PLoS Biology*. **2(9)**, e268.

Falush D, Stephens M, Pritchard JK. (2003). "Inference of populatiuon structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics*. **164**, 1567-1587.

Felsenfeld G, Davis DR, Rich A. (1957). "Formation of a three-stranded polynucleotide molecule". *J. Am. Chem. Soc*. **79**, 2023–2024.

Fenech AG, Ebejer MJ, Felice AE, Ellul-Micallef R, Hall IP. (2001). "Mutation screening of the muscarinic M(2) and M(3) receptor genes in normal and asthmatic subjects". *Br. J. Pharmacol*. **133**, 43-48.

Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. (2004). "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins". *Nat Biotechnol.* **22**, 1302-1306.

Ferrer-Costa C, Orozco M, de la Cruz X. (2002). "Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties". *J. Mol. Biol*. **315**, 771–786.

Ferrer-Costa C, Orozco M, de la Cruz X. (2004). "Sequence-based prediction of pathological mutations". *Proteins.* **57**, 811–819.

Ferrer-Costa C, Orozco M, de la Cruz X. (2005). "Use of bioinformatics tools for the annotation of disease-associated mutations in animal models". *Proteins.* **61**, 878–887.

Fickett JW, Wasserman WW. (2000). "Discovery and modeling of transcriptional regulatory regions". *Current Opinion in Biotechnology*. **11**, 19-24.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol, EJ, Smoller JW, Pato CN, *et al*. (2004). "Assessing the impact of population stratification on genetic association studies". *Nat. Genet*. **36**, 388-393.

Gabellini N. (2001). "A polymorphic GT repeat from the human cardiac Na+Ca2+ exchanger intron 2 activates splicing". *Eur J Biochem*, **268(4)**, 1076–83.

Gershenzon NI, Ioshikhes IP. (2005). "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis". *Bioinformatics*. **21**, 1295–1300.

Glazier AM, Nadeau JH, Aitman TJ. (2002). "Finding genes that underlie complex traits". *Science*. **298**, 2345–2349.

Goddard KA, Hopkins PJ, Hall JM, Witte JS. (2000). "Linkage disequilibrium and allele frequency distributions for 114 single-nucleotide polymorphisms in five populations". *Am J Hum Genet*. **66**, 216-34.

Goñi JR, de la Cruz X, Orozco M. (2004). "Triplex-forming oligonucleotide target sequences in the human genome". *Nucleic Acids Res*. **32**, 354-60.

Gosens R, Zaagsma J, Meurs H, Halayko AJ. (2006). "Muscarinic receptor signaling in the pathophysiology of asthma and COPD". *Respir Res.* **7**, 73.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. (2006). "miRBase: microRNA sequences, targets and gene nomenclature". *Nucleic Acids Res*. **34**, D140-D144.

GuhaThakurta D, Xie1 T, Anand M, Edwards SW, Li G, Wang SS, Schadt EE. (2006). "Cis-regulatory variations: A study of SNPs around genes showing cis-linkage in segregating mouse populations". *BMC Genomics*. **7**, 235.

Guigó R. (1998). "Assembling genes from. predicted exons in linear time with dynamic programming". *J. Comput. Biol*. **5**, 681–702.

Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. (1999). "O-GLYCBASE version 4.0: a revised

database of O-glycosylated proteins". *Nucleic Acids Res*. **27**, 370–372

Hartl DL, Clark AG. (1997). "Principles of Population Genetics". 3ªed. Sinauer Associates, Sunderland, MA. 519 pp

Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der A DL, Feskens EJ. (2006). "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases". *BMC Genetics*. **7**, 23.

Herrero J, Al-Shahrour F, Diaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J. (2003). "GEPAS: a web-based resource for microarray gene expression data analysis". *Nucleic Acids Res*. **31**, 3461-3467

Hirschhorn JN, Daly MJ. (2005)."Genome-wide association studies for common diseases and complex traits". *Nat Rev Genet.* **6**, 95–108.

Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J. (2000). "Selecting SNPs in two-stage analysis of disease association data: a model-free approach". *Ann Hum Genet*. **64**, 413–417.

Hoh J, Wille A, Ott J. (2001). "Trimming, weighting, and grouping SNPs in human case-control association studies". *Genome Res*. **11**, 2115–2119.

Hoh J, Ott J. (2003). "Mathematical multi-locus approaches to localizing complex human trait genes". *Nat Rev Genet*. **4(9)**, 701-709.

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*. (2007). "Ensembl 2007". *Nucleic Acids Res*. **35**, D610-7.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. (2003). "Widespread purifying selection at polymorphic sites in human protein-coding loci". *Proc. Natl. Acad. Sci. USA*. **100**, 15754–15757.

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, *et al*. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease". *Nature*. **411**, 599–603.

Hupé P, Stransky S, Thiery JP, Radvanyi F, Barillot E. (2004) "Analysis of array CGH data: from signal ratio to gain and loss of DNA regions". *Bioinformatics*. **20**, 3413-22.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. (2004). "Detection of large-scale variation in the human genome". *Nat Genet*. **36**, 949-51.

International HapMap Consortium (2005). "A haplotype map of the human genome". *Nature*. **437**, 1299-1320

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. (2001). "Replication validity of genetic association studies". *Nat. Genet*. **29**, 306-309.

Ishii S, Nakao S, Minamikawa-Tachino R, Desnick RJ, Fan JQ. (2002). "Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype". *Am J Hum Genet*. **70(4)**, 994-1002.

Jegga AG, Gowrisankar S, Chen J, Aronow BJ. (2007). "PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease". *Nucl. Acids Res*. **35**, D700-D706.

Joo J, Tian X, Zheng G, Lin JP, Geller NL. (2005). "Selection of single-nucleotide polymorphisms in disease association data". *BMC Genetics*. **6**, S93.

Julia A, Moore J, Miquel L, Alegre C, Barcelo P, Ritchie M, Marsal S. (2007). "Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction". *Genomics*. **90(1)**, 6-13.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. (1992). "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors". *Science*. **258**, 818–821.

Kanhere A, Bansal M. (2005). "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes". *Nucleic Acids Res*. **33**, 3165–3175.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. (2005). "LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources". *Bioinformatics*. **21**, 2814-2820.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. (2003). "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences". *Nucleic Acids Res*. **31**, 3576–3579.

Kim SY, Nam SW, Lee SH, Park WS, Yoo NJ, Lee JY, Chung YJ. (2005). "ArrayCyGHt: a web application for analysis and visualization of array-CGH data". *Bioinformatics*. **21**, 2554–2555.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. (2007). "A randomization test for controlling population stratification in whole-genome association studies". *Am J Hum Genet. In press.*

Knight JC. (2005). "Regulatory polymorphisms underlying complex disease traits". *J. Mol. Med*. **83**, 97–109.

Kreegipuu A, Blom N, Brunak S. (1999). "PhosphoBase, a database of phosphorylation sites: release 2.0.". *Nucleic Acids Res*. **29**, 237–239.

Krishnan VG, Westhead DR. (2003). "A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function". *Bioinformatics*. **19(17)**, 2199–2209

Kruglyak L. (1999). "Prospects for whole-genome linkage disequilibrium mapping of common disease genes". *Nat Genet*. **22**, 139-44.

Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, *et al*. (2007). "The UCSC genome browser database: update 2007". *Nucleic Acids Res*. **35**, D668-73.

Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB". Genes Dev. 12(1), 34-44.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*. (2001). "Initial sequencing and analysis of the human genome". *Nature.* **409**, 860-921.

Larsen F, Gundersen G, Lopez R, Prydz H. (1992). "CpG islands as gene markers in the human genome". *Genomics*. **13(4)**, 1095-107.

Le Bellego F, Plante S, Chakir J, Hamid Q, Ludwig MS. (2006). "Differences in MAP Kinase Phosphorylation in Response to Mechanical Strain in Asthmatic Fibroblasts". *Respir Res*. **7(1)**, 68.

Lee C. (2005). "Vive la difference!". *Nature Genet.* **37**, 660–661.

Lee JA, Lupski JR. (2006). "Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders". *Neuron*. **52**, 103-21.

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. (2003). "Identification of conserved regulatory elements by comparative genome analysis". *J Biol*. **2(2)**, 13.

Liva S, Hupe P, Neuvial P, Brito I, Viara E, La Rosa P, Barillot E. (2006). "CAPweb: a bioinformatics CGH array analysis platform". *Nucleic Acids Res*. **34**, W477-W481.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. (2003). "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease". *Nat. Genet.* **33**, 177-182.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons". *Science*. **288**, 136-140.

López KI, Martínez SE, Moguel MC, Romero LT, Figueroa CS, Pacheco GV, Ibarra B, Corona JS. (2007). "Genetic diversity of the IL-4, IL-4 receptor and IL-13 loci in mestizos in the general population and in patients with asthma from three subpopulations in Mexico". *Int J Immunogenet.* **34(1)**, 27-33.

López de la Paz M, Serrano L. (2004). "Sequence determinants of amyloid fibril formation". *Proc. Natl Acad. Sci. USA*. **101**, 87–92.

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. (2004). "Screening large-scale association study data: exploiting interactions using random forests". *BMC Genet*. **5**, 32.

Lupski JR. (2007). "Structural variation in the human genome". *N Engl J Med*. **356(11)**, 1169-71.

Lynch KW, Weiss A. (2001). "A CD45 Polymorphism Associated with Multiple Sclerosis Disrupts an Exonic Splicing Silencer". *J. Biol. Chem*. **276**, 24341-24347.

Maquat LE. (2004). "Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics". *Nat. Rev. Mol. Cell Biol*. **5**, 89–99.

Massingham T, Goldman N. (2005). "Detecting amino acid sites under positive selection and

purifying selection". *Genetics*. **169**, 1753–1762.

McVety S, Li L, Gordon PH, Chong G, Foulkes WD. (2006). "Disruption of an exon splicing enhancer in exon 3 of MLH1 is the cause of HNPCC in a Quebec family". *J Med Genet*. **43(2)**, 153-6.

Møller LB, Tümer Z, Lund C, Petersen C, Cole T, Hanusch R, Seidel J, Jensen LR, Horn N. (2000). "Similar splice-site mutations of the ATP7A gene lead to different phenotypes: Classical Menkes disease or occipital horn syndrome". *Am J Hum Genet.* **66,** 1211-1220.

Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, *et al*. (2006). "Next station in microarray data analysis: GEPAS". *Nucleic Acids Res.* **34**, W486–W491.

Montgomery SB, Griffith OL, JSchuetz JM, Brooks-Wilson A, Jones SJM. (2007). "A Survey of Genomic Properties for the Detection of Regulatory Polymorphisms". PLoS Comput Biol. 3(6), e106.

Mooney S. (2005). "Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis". *Brief. Bioinform*. **6**, 44-56.

Mooney SD, Klein TE, Altman RB, Trifiro MA, Gottlieb B. (2003). "A functional analysis of disease-associated mutations in the androgen receptor gene". *Nucleic Acids Res.* **31(8)**, e42.

Moore JH, Lamb JM, Brown NJ, Vaughan DE. (2002). "A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE i/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels". *Clin Genet.* **62**, 74–79.

Moore JH, Williams SM. (2002). "New strategies for identifying genegene interactions in hypertension". *Ann Med.* **34**, 88–95.

Motsinger AA, Lee SL, Mellick G, Ritchie MD. (2006). "GPNN: Power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease". *BMC Bioinformatics.* **7**, 39.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. (1986). "Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction". *Cold Spring Harbor Symposia on Quantitative Biology.* **51(Pt 1)**, 263–273.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. (1992). "Specific enzymatic amplification

of DNA in vitro: the polymerase chain reaction". *Biotechnology.* **24**, 17–27.

Nagelkerke N, Smits J, Le Cessie S, Van Houwelingen H. (2005). "Testing goodness-of-fit of the logistic regression model in case-control studies using sample reweighting". *Statist Med.* **24**, 121-130.

Nakai K, Horton P. (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization". *TIBS.* **24**, 34–35.

Nelson MR, Kardia SL, Ferrell RE, Sing CF. (2001). "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation". *Genome Res.* **11,** 458–470.

Newton-Cheh C, Hirschhorn JN. (2005). "Genetic association studies of complex traits: design and analysis issues". *Mutation Research.* **573**, 54-69.

Ng PC, Henikoff S. (2001). "Predicting deleterious amino acid substitutions". *Genome Res.* **11(5)**, 863–874.

Ng PC, Henikoff S. (2003). "SIFT: Predicting amino acid changes that affect protein function". *Nucleic Acids Res.* **31(13)**, 3812–3814.

Ng PC, Henikoff S. (2006). "Predicting the Effects of Amino Acid Substitutions on Protein Function". *Annual Review of Genomics and Human Genetics*. **7**, 61-80.

Nishikawa J, Amano M, Fukue Y, Tanaka S, Kishi H, Hirota Y, Yoda K, Ohyama T. (2003). "Left-handedly curved DNA regulates accessibility to cis-DNA elements in chromatin". *Nucleic Acids Res.* **31**, 6651-6662.

Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data". *Biostatistics.* **5(4)**, 557-572.

Pagani F, Stuani C, Tzetis M, Kanavakis E, Efthymiadou A, Doudounakis S, Casals T, Baralle FE. (2003). "New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12". *Hum. Mol. Genet.* 1**2**, 1111–1120.

Pauling L, Corey RB. (1953). "A proposed structure for the nucleic acids". *Proc. Natl Acad. Sci. USA.* **39**, 84-97.

Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW. (2000). "A DNA structural atlas for

Escherichia coli". *J. Mol. Biol*. **299**, 907– 930.

Plagnol V, Cooper JD, Todd JA, Clayton DG. (2007). "A Method to Address Differential Bias in Genotyping in Large-Scale Association Studies". *PLoS Genetics*. **3(5)**, e74.

Polzehl J, Spokony S. (2000). "Adaptative weights smoothing with applications to image restoration". *J. R. Stat. Soc. Ser. B*. **62**, 335-354.

Prestridge DS. (1995). "Predicting Pol II promoter sequences using transcription factor binding sites". *J. Mol. Biol*. **249**, 923–932.

Pritchard JK, Rosenberg NA. (1999). "Use of unlinked genetic markers to detect population stratification in association studies". *Am J Hum Genet.* **65**, 220-28.

Pritchard JK, Stephens M, Donnelly P. (2000a). "Inference of population structure using multilocus genotype data". *Genetics*. **155(2)**, 945-59.

Pritchard JK, Stephens M, Rosenberg NA, Donelly P. (2000b). "Association mapping in structured populations". *Am J Hum Genet.* **67**, 170-181.

Pritchard JK. (2001). "Are rare variants responsible for susceptibility to complex diseases?". *Am J Hum Gen.* **69**, 124-137.

Qin ZS, Niu T, Liu JS. (2002). "Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms". *Am J Hum Genet*. **71(5)**, 1242-7.

Quilliam LA, Zhong S, Rabun KM, Carpenter JW, South TL, Der CJ, Campbell-Burk S. (1995). "Biological and structural characterization of a Ras transforming mutation at the phenylalanine-156 residue, which is conserved in all members of the Ras superfamily". *Proc. Natl Acad. Sci. USA*. **92(5)**, 1272–1276.

Ramensky V, Bork P, Sunyaev S. (2002). "Human non-synonymous SNPs: Server and survey". *Nucleic Acids Res*. **30(17)**, 3894–3900.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, *et al*. (2006). "Global variation in copy number in the human genome". *Nature.* **444**, 444-54.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. (2001). "Linkage disequilibrium in the human genome". *Nature*. **411**, 199-204.

Reich DE, Lander ES. (2001). "On the allelic spectrum of human disease". *Trends Genet*. **17**, 502-510.

Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F. (2005). "SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs". *Nucleic Acids Res.* **33**, D527-D532.

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. (2006). "SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs". *Bioinformatics*. **22(17)**, 2183-5.

Risch N, Merikangas K. (1996). "The future of genetic studies of complex human diseases". *Science*. **273**, 1516-7.

Risch NJ. (2000). "Searching for genetic determinants in the new millennium". *Nature*. **405(6788)**, 847–856.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. (2001). "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer". *Am J Hum Genet*. **69**, 138–147.

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. (2003a). "Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases". *BMC Bioinformatics*. **4**, 28.

Ritchie MD, Hahn LW, Moore JH. (2003b). "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity". *Genet Epidemiol.* **24**, 150–157.

Rosenberg NA, Li LM, Ward R, Pritchard JK. (2003). "Informativness of genetic markers for inference of ancestry". *Am. J. Hum. Genet*. **73**, 1402-1422.

Rudd MF, Williams RD, Webb EL, Schmidt S, Sellick GS, Houlston RS. (2005). "The predicted impact of coding single nucleotide polymorphisms database". *Cancer Epidemiol Biomarkers Prev.* **14(11)**, 2598-604.

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles". *Nucleic Acids Res*. **32**, D91-94.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. (2007). "Challenges and standards in integrating surveys of structural variation". *Nature Genetics.* **39**, S7-S15.

Scherf M, Klingenhoff A, Werner T. (2000). "Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach". *J. Mol. Biol.* **297(3)**, 599-606.

Schmid CD, Perier R, Praz V, Bucher P. (2006). "EPD in its twentieth year: towards complete promoter coverage of selected model organisms". *Nucleic Acids Res*. **34**, D82-5.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. (2003). "Human-Mouse Alignments with BLASTZ". *Genome Res*. **13**, 103-107.

Schwender H, Zucknick M, Ickstadt K, Bolt HM. (2004). "A pilot study on the application of statistical classification procedures to molecular epidemiological data". *Toxicol Lett*. **151**, 291–299.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. (2005). "The FoldX web server: an online force field". *Nucleic Acids Res*. **33**, W382-388.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M. (2004). "Large-scale copy number polymorphism in the human genome". *Science*. **305**, 525- 8.

Sebat J. (2007). "Major changes in our DNA lead to major changes in our thinking". *Nature Genetics.* **39**, S3-S5.

Sham PC, Curtis D. (1995). "An extended transmission/disequilibrium test (TDT) for multi-allele marker loci". *Annals of Human Genetics*. **59**, 323–336.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, *et al.* (2005) "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". *Genome Res*. **15**, 1034-1050.

Sironi M, Menozzi G, Riva L, Cagliani R, Comi GP, Bresolin N, Giorda R, Pozzoli U. (2004). "Silencer elements as possible inhibitors of pseudoexon splicing". *Nucleic Acids Res*. **32**, 1783–1791.

Slatkin M. (1994). "Linkage disequilibrium in growing and stable populations". *Genetics*. **137**, 331-336.

Smale ST, Baltimore D. (1989). "The "initiator" as a transcription control element". *Cell.* 57, 103–113.

Smale ST, Kadonaga JT. (2003). "The RNA polymerase II core promoter". *Annu. Rev. Biochem*. **72**, 449–479.

Sousa AR, Lane SJ, Soh C, Lee TH. (1999). "In vivo resistance to corticosteroids in bronchial asthma is associated with enhanced phosyphorylation of JUN N-terminal kinase and failure of prednisolone to inhibit JUN N-terminal kinase phosphorylation". *J Allergy Clin Immunol*. **104**, 565-74.

Souverein OW, Zwinderman AH, Tanck MW. (2006). "Multiple imputation of missing genotype data for unrelated individuals". *Ann Hum Genet*. **70**, 372-81.

Spielman RS, McGinnis RE, Ewens WJ. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)". *American Journal of Human Genetics*. **52(3)**, 506–516.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. (2006). "BioGRID: a general repository for interaction datasets". *Nucleic Acids Res*. **34**, D535-9.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2007). "Human Gene Mutation Database (HGMD®): 2003 update". *Human Mutation.* **21(6)**, 577-581.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. (2004). "topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association". *Nucleic Acids Res*. **32**, D520–D522.

Stormo GD. (2000). "DNA binding sites: representation and discovery". *Bioinformatics.* **16**, 16-23.

Strittmatter WJ, Roses AD. (1996). "Apolipoprotein E and Alzheimer's disease". *Annu Rev Neurosci.* **19**, 53-77.

Sun F, Flanders WD, Yang Q, Khoury MJ. (1999). "Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT". *American Journal of Epidemiology*. **150**, 97–104.

The Wellcome Trust Case Control Consortium. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". *Nature*. **447**, 661-78.

Thompson MD, Takasaki J, Capra V, Rovati GE, Siminovitch KA, Burnham WM, Hudson TJ, Bossé Y, Cole DE. (2006). "G-protein-coupled receptors and asthma endophenotypes: the cysteinyl leukotriene system in perspective". *Mol Diagn Ther*. **10(6)**, 353-66.

Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H. (2004). "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma". *Bioinformatics*. **5**, 120.

Treisman R , Orkin SH, Maniatis T. (1983). "Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes". *Nature*. **302**, 591–596.

Tsai L, Luo L, Sun Z. (2002). "Sequence-dependent flexibility in promoter sequences". *J. Biomol. Struct. Dyn*. **20**, 127– 134.

Tuerk C, Gold L. (1990). "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase". *Science*, **249**, 505–510.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*. (2001). "The sequence of the human genome". *Science*. **291**, 1304-51.

Vercelli D. (2003). "Learning from discrepancies: CD14 polymorphisms, atopy and the endotoxin switch". *Clin Exp Allergy*. **33**, 153-155.

Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. (2006). "SNP Function Portal: a web database for exploring the function implication of SNP alleles". *Bioinformatics*. **22(14)**, e523-e529.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. (2004). "Systematic identification and analysis of exonic splicing enhancers". *Cell*, **119**, 831-845.

Wasserman WW, Sandelin A. (2004). "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics*, **5**, 276-287.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*. (2002). "Initial sequencing and comparative analysis of the mouse genome". *Nature*. **420**, 520-562.

Weber JL, May PE. (1989). "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction". *American Journal of Human Genetics*. **44(3)**, 388–396.

Weiss KM, Clark AG. (2002). "Linkage disequilibrium and the mapping of complex human traits". *Trends Genet*. **18**, 19-24.

Wieringa B, Meyer F, Reiser J, Weissmann C. (1983). "Unusual splice sites revealed by mutagenic inactivation of an authentic splice site of the rabbit beta-globin gene". *Nature*. **301**, 38–43.

Wille A, Hoh J, Ott J. (2003). "Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers". *Genet Epidemiol*. **25**, 350-359.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M., Reuter I, Schacherer F. (2000). "TRANSFAC: an integrated system for gene expression regulation". *Nucleic Acids Res*. **28**, 316–319.

Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Züchner S, Hauser MA. (2005). "SNPselector: a web tool for selecting SNPs for genetic association studies". *Bioinformatics*. **21(22)**, 4181-6.

Yang Z. (1997). "PAML: A program package for phylogenetic analysis by maximum likelihood". *Comput. Appl Biosci*. **13**, 555–556.

Yang Z, Nielsen R. (2002). "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages". *Mol. Biol. Evol*. **19**, 908–917.

Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN. (2006). "FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization". *Nucleic Acids Res*. **34**, W635-41.

Zang Y, Zhang H, Yang Y, Zheng G. (2007). "Robust genomic control and robust delta centralization test for case-control association studies". *Human heredity*. **63(3-4)**, 187-95.

Zeggini E, Rayner W, Morris AP, Hattersley AT, Walter M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI. (2005). "An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets". *Nat. Genet*. **37**, 1320-1322.

Zhang L, Vincent GM, Baralle M, Baralle FE, Anson BD, Benson DW, Whiting B, Timothy KW,

Carlquist J, January CT, *et al*. (2004). "An intronic mutation causes long QT syndrome". *J. Am. Coll. Cardiol*. **44**, 1283–1291.

Zhang XH, Chasin LA. (2004). "Computational definition of sequence motifs governing constitutive exon splicing". *Genes Dev*. **18**, 1241–1250.

Zhang Z, Gerstein M. (2003). "Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements". *Journal of Biology*. **2**, 11.

Zhao Z, Fu Y-X, Hewett-Emmett D, Boerwinkle E. (2003). "Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution". *Gene*. **312**, 207–213.

Zimmermann N, King NE, Laporte J, Yang M, Mishra A, Pope SM, Muntel EE, Witte DP, Pegg AA, Foster PS, *et al*. (2003). "Dissection of experimental asthma with DNA microarray analysis identifies arginase in asthma pathogenesis". *J. Clin. Invest.* **111**, 1863–1874.

ANEXO

PUBLICACIONES

Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, Robledo M, Dopazo J. (2004). "PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level". *Nucleic Acids. Res*. 32, W242-W248.

Conde L, Vaquerizas JM, Ferrer-Costa C, Orozco M, Dopazo J. (2005). "PupasView: a visual tool for selecting suitable SNPs, with putative pathologic effect in genes, for genotyping purposes". *Nucleic Acids Res*. **33**, W501-5.

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J. (2006). "PupaSuite: finding functional SNPs for large-scale genotyping purposes". *Nucleic Acids Res.* **34,** W621-W625.

Conde L, Montaner D, Burguet-Castell J, Tarraga J, Medina I, Al-Shahrour F, Dopazo J. (2007). "ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling". *Nucleic Acids Res.* **35**, W81–W85.

Conde L, Montaner D, Burguet-Castell J, Tarraga J, Al-Shahrour F, Dopazo J. (2007). "Functional profiling and gene expression analysis of chromosomal copy number alterations". *Bioinformation*. **1(10)**, 432-435.

W242–W248 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh438

PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level

Lucía Conde, Juan M. Vaquerizas, Javier Santoyo, Fátima Al-Shahrour, Sergio Ruiz-Llorente[1], Mercedes Robledo[1] and Joaquín Dopazo*

Bioinformatics Unit and [1]Hereditary Endocrine Cancer Group, Centro Nacional de Investigaciones Oncológicas (CNIO) Madrid, Spain

Received February 3, 2004; Revised and Accepted April 15, 2004

ABSTRACT

We have developed a web tool, PupaSNP Finder (PupaSNP for short), for high-throughput searching for single nucleotide polymorphisms (SNPs) with potential phenotypic effect. PupaSNP takes as its input lists of genes (or generates them from chromosomal coordinates) and retrieves SNPs that could affect the conserved regions that the cellular machinery uses for the correct processing of genes (intron/exon boundaries or exonic splicing enhancers), predicted transcription factor binding sites (TFBS) and changes in amino acids in the proteins. The program uses the mapping of SNPs in the genome provided by Ensembl. Additionally, user-defined SNPs (not yet mapped in the genome) can be easily provided to the program. Also, additional functional information from Gene Ontology, OMIM and homologies in other model organisms is provided. In contrast to other programs already available, which focus only on SNPs with possible effect in the protein, PupaSNP includes SNPs with possible transcriptional effect. PupaSNP will be of significant help in studies of multifactorial disorders, where the use of functional SNPs will increase the sensitivity of identification of the genes responsible for the disease. The PupaSNP web interface is accessible through http://pupasnp.bioinfo.cnio.es.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and they represent one of the most powerful tools for the analysis of genomes (1). Owing to their widespread distribution, SNPs are particularly valuable as genetic markers in the search for disease susceptibility genes, drug response-determining genes, and so on. In the past decades, linkage analysis has been very successful in the identification of genes responsible for mendelian diseases. Nevertheless, direct application of linkage analysis to the case of complex diseases, in which several genes with weaker genotype–phenotype correlations are involved, has resulted in more modest success (2). Now, it is believed that improved genotyping methods in combination with the proper design strategies could bring the genetics of complex diseases to a point of success comparable to where mendelian genetics now firmly resides (3).

There are examples documented in which alleles of more than one gene contribute to the same disease. It is generally believed that multigenic diseases reflect disruptions in the proteins that participate in a protein complex or a pathway (4). Typically, SNPs have been used as markers; that is, the real determinant of the disease was not the SNP itself but some other mutation in linkage disequilibria with it.

The use of functional SNPs could be an important factor for increasing significantly the sensitivity of association tests. In fact, several complex genetic disorders such as Alzheimer's disease (5) and Crohn's disease (6) have been associated with functional SNPs, lending credence to strategies giving priority to candidate markers based on predictable function. The latest build of NCBI's dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi) contains 5 772 564 SNPs, with 2 356 957 of them validated. This means that human variation has been screened to an average resolution of 1 SNP for every 566 nt. There is also curated information on SNPs in HGVbase (7). These figures suggest that the possibility of finding the real determinant of a disease among the characterized SNPs can be seriously considered. In fact, dbSNP build 117 contains 24 483 SNPs located in coding regions that produce amino acid change, affecting a total of 9791 different genes. Several estimate suggest that, overall, only 20% of them could damage

*To whom correspondence should be addressed. Tel: +34 912246919; Fax: +34 912246972; Email: jdopazo@cnio.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

the protein (8). Much attention has been focused on the possible phenotypic effects of SNPs that cause amino acid changes. The volume of available information together with the development of more sophisticated methods of protein structure prediction has led to different attempts to relate the effect of amino acid changes to structural distortions and, consequently, possible phenotypic effect. Following this, two main different approaches have been taken: on the one hand is the study of conservation of residues in homologous proteins (9) including more sophisticated approaches taking into account the phylogenetic history (10) and, on the other hand, there is the study of changes in the stability (11,12) and other properties of the protein due to changes of amino acids (8,13).

Nevertheless, there are different ways in which the functionality of a gene product can be affected without requiring a amino acid change in the protein. There is increasing evidence that many human disease genes harbour exonic or non-coding mutations that affect pre-mRNA splicing (14). Alternative splicing produced by mutations in intron/exon junctions, or in distinct binding motifs, such as exonic splicing enhancers (ESEs), to which different proteins involved in splicing bind, is the basis of different diseases. In fact, it has been estimated that 15% of point mutations that result in human genetic diseases cause RNA splicing defects (15). For example, a silent mutation in exon 14 of the *APC* gene is associated with exon skipping in a Familial Adenomatous Polyposis (FAP) family (16), and there are many more examples [see Table 2 in (14)]. Also, alterations in the level of expression of gene products can cause diseases. Different SNPs are associated with alterations in gene expression (17) and, in some cases, it is known that they alter some regulatory sequence motif. For example, a regulatory polymorphism in the programmed cell death 1 gene (*PDCD1*), which alters a binding site for the runt-related transcription factor 1 (*RUNX1*) located in an intronic enhancer, is associated with susceptibility to systemic lupus erythematosus in humans (18). It has also been reported that polymorphisms in the gelatinase A promoter region are associated with diminished transcriptional response to estrogen and genetic fitness (19). A recent large-scale screening over a set of 16 chromosomes, found SNPs in the promoters regions of 35% of the genes, and experimental evidence suggested that around one-third of promoter variants may alter gene expression to a functionally relevant extent (20). Therefore, the inclusion of other possible causes of loss of functionality in gene products, beyond the simple estimation of the possible phenotypic effect of an amino acid change, increases considerably the number of SNPs with potential phenotypic effect to be considered for the design of experiments.

Classical statistical linkage tests need a large number of cases if the number of genes to be tested is high. It has only recently been recognized that reliable identification of genetic variants that affect gene regulation is still a challenge in genomics and is expected to play an important role in the molecular characterization of complex traits (21). Another important consideration when analysing multigenic traits is the information available on the genes. Information allows a more targeted approach, by focusing initially on genes whose functionality is related to the disease studied.

Genome surveys based on the information contained in dbSNP show that there are 361 SNPs mapped in splice sites of introns, 1 387 506 in introns and 242 842 in untranslated regions affecting 336 16 306 and 14 198 genes, respectively. A number of these SNPs could be disease determinants.

With the idea of extracting as much information as possible form SNPs with putative phenotypic effect, we have developed PupaSNP Finder (Putative Phenotypic Alterations caused by SNPs; PupaSNP for short). This tool retrieves all the SNPs present in a set of genes of interest that potentially affect the functionality of the gene product. This list is combined with functional information obtained from Gene Ontology (GO) annotations (22). Genes can be directly retrieved from genomic locations or, alternatively, can be taken from a list provided by the user. This corresponds to two typical problems: (i) traits mapped to a given chromosomal region or (ii) traits associated with a given class of genes (e.g. a signalling pathway). Genome coordinates of genes and SNPs are taken from the Ensembl annotation (23).

METHODS

Finding SNPs with potential phenotypic effect

PupaSNP operates with a collection of entries from dbSNP mapped to the Golden Path genome assembly, as implemented in human section of Ensembl (http://www.ensembl.org). As previously mentioned, PupaSNP uses a list of genes and generates a report in which all the SNPs with possible phenotypic effect are listed. The genes can be selected directly by their location in a region of the genome, or just provided as a list (e.g. genes belonging to a given pathway, involved in a particular biological function). Genomic regions can be selected either by defining a range of chromosome coordinates or by directly choosing the cytoband of interest. The engine finds all the genes located within the specified region as well as their promoter regions using Ensembl APIs. In the case of a user-defined list, Ensembl is used to extract their complete intron/exon structure as well as the promoter regions.

The potential effects on the phenotype taken into account are at both transcriptional and gene product levels. These include alterations in (i) transcription factor binding sites, (ii) intron/exon border consensus sequences, (iii) ESE sequences, which are the binding sites for specific serine/arginine-rich (SR) proteins involved in the splicing machinery (24,25) and (iv) the exons that cause an amino acid change. Additionally, the GO terms (22) associated with the genes can be obtained. This is very useful in the case of looking for genes in a chromosomal region, because it can help to discard genes definitively not involved in the disease studied, based on the annotations.

Transcription factor binding sites. In the search for SNPs with potential phenotypic effect, 10 000 bp upstream of the genes, belonging to the promoter region of each gene in the list, are scanned for the presence of possible transcription factor binding sites (TFBSs). The program Match™ (26), version 1.10, from the Transfac® database (27), version professional 7.3, was used for this purpose. SNPs located within these motifs are considered to have a putative phenotypic effect in the expression of the gene. The options used for the program Match™ were (i) group of matrices: vertebrates, (ii) use high quality matrices only and (iii) cutoff selection for

matrix group: to minimize false positives. This cutoff was obtained by exploring the third exon sequences with the weight matrices and was chosen to reduce the number of random putative sites found by the program (26).

Although the scan is done in a region 10 000 bp upstream from the start of the gene, the number of bases to be taken into account in the study is customizable. Obviously, the closer to the start of the gene, the more likely the binding site is to be authentic.

Intron–exon boundaries. Ensembl APIs were used to extract the intron/exon organization of the genes and the corresponding sequences. The two conserved nucleotides at each side of the splicing point, which constitute the splicing signal (14), were then located and all the SNPs altering these signals are recorded.

Exonic splicing enhancers. Mutations that deactivate or activate exonic splicing enhancer sequences may result in exon skipping, malformation, and so on. ESEs also appear to be important in exons that normally undergo alternative splicing. Different classes of ESE consensus motifs have been described, but they are not always easily identified. We have developed a script that scans exon sequences to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, by using the weight matrices available for them (28). A score is obtained related to the likelihood that the site found is a real ESE. Only ESE sites with scores over the threshold [see (28) and http://exon.cshl.org/ESE/ESEmatrix.html for details] are taken into account in the analysis. Threshold values, above which a score for a given sequence is considered to be significant, are set as the median of the highest score for each sequence in a set of 30 randomly chosen 20 nt sequences (from the starting pool used for functional assays for ESE identification; see http://exon.cshl.org/ESE/ESEmatrix.html). If an SNP disrupts one of these sequences, the new score, corresponding to the mutated sequence, is also calculated. Strong differences between the two score values suggest more drastic effects caused by the SNP.

Changes at amino acid level and functional implications. SNPs that result in a change of amino acid are likely to cause some phenotypic effect and, consequently, are all listed. Since the main purpose of the tool is to cover possible transcriptional effects of the SNPs and there are a number of tools already available for the prediction of phenotypic effects due to mutations in amino acids (see Introduction) PupaSNP only lists them. To help in the identification of possible effects we label SNPs that disrupt any functional motif as listed in Interpro (29), a resource that compiles information on protein families, domains and functional sites. The coordinates of the Interpro motifs within the exons of the genes are extracted from Ensembl and cross-referenced with the SNPs coordinates.

Additional functional information. Since PupaSNP finder works with lists of genes in order to select the best SNP candidates for further use in association analysis, it is very helpful to have functional annotations of the genes. This allows the assignment of priorities based also on the information available on the genes. Information is obtained from (i) Gene Ontology annotations, obtained through the FatiGO engine (30) (available at http://fatigo.bioinfo.cnio.es), (ii) OMIM (Online Mendelian Inheritance in Man), which constitutes a comprehensive, authoritative and timely knowledge base of human genes and genetic disorders (31) and (iii) homologies to other organisms, obtained directly from Ensembl. Gene Ontology is a tree structure (called a directed acyclic graph) in which terms describing three fundamental ontologies (molecular function, biological process and cellular component) have descendants with more detailed descriptions. Thus, descending the hierarchy of GO implies moving towards terms with more detailed descriptions of the ontologies, but, at the same time, there are fewer genes with annotations at such detail. FatiGO works by climbing up the hierarchy to a selected parent level (30) to optimize the number of genes with annotation and the detail of the annotation. Thus, the identification of common parent functions or processes is easier. In this way, the consideration of the SNPs in a functional context can help to understand the potential biological implications of the SNPs and genes studied.

RESULTS

SNPs with possible phenotypic effect

We analysed a total of 24 037 human genes corresponding to the annotations in Ensembl build 34 (version 18.34.1), which contains the mapping of dbSNP 117. By scanning with the Match^{TM} program the 10 000 bp upstream promoter regions of the genes, 2 587 478 transcription factor binding sites, corresponding to 330 different Transfac weight matrices (27), were found. After mapping the SNPs in the promoter regions, 71 444 TFBSs were found to be disrupted by a total of 57 412 SNPs (some SNPs affect more than one TFBS at the same time). A total of 19 010 genes presented at least 1 predicted TFBS disrupted by a SNP, which constitutes a considerable proportion of the total number of genes. The coverage in terms of both SNPs and TFBS predictions was good: only for 54 genes was no single SNP found in the 10 000 bp 5′-upstream region, and only for 2 genes could no predicted TFBS be found (*ENSG00000116119*, or *KV2A_HUMAN*, which is the IG KAPPA CHAIN V-II REGION CUM, and *ENSG00000174994*, or *AK057375*, which seems to be a DNA binding protein). In a number of cases, SNPs affect overlapping TFBSs, which could have a stronger effect still in the phenotype. There are even 2 SNPs that simultaneously affect 15 TFBSs.

The four conserved bases that define intron–exon boundaries were mutated by 844 SNPs, affecting to a total of 598 genes.

Over eight million ESE motifs were found, covering all the genes studied. A total of 138 746 SNPs were found to disrupt ESE sequences. These SNPs affect a total of 17 312 genes.

These results suggest that, in the search for SNPs with potential phenotypic effects, regulatory SNPs or SNPs affecting splicing should not be neglected.

The web interface

Input data. PupaSNP has been designed for high-throughput screening of functional SNPs. Thus, the input consists of a list of genes. The list can be directly provided as a collection of gene identifiers (Ensembl IDs, or external IDs, which include

GenBank, Swissprot/TrEMBL and other gene IDs supported by Ensembl) or can be specified by means of a chromosomal location (cytobands or chromosomal coordinates). In the latter case, PupaSNP extracts all the genes contained in the specified location. Ensembl coordinates are used to extract the genes. Only Ensembl annotated genes, but not predictions, are extracted.

User-defined SNPs. Alternatively, the user can input SNPs not in the database in a very straightforward manner and take advantage of the tools for predicting their potential phenotypic effect. A text file containing the descriptions of the SNPs must be generated. Each line describes one unique SNP with the following tab-delimited data: SNP name, gene (Ensembl ID or external ID), position with respect to the start of the translation and alleles, e.g.

```
MySNP01  ENSG00000000003  -1830  A/G
MySNP02  ENSG00000157873  421    C/G
```

This describes two SNPs: the first in the gene *ENSG00000000003* (*tetraspanin 6*, or *TSPAN6*), 1830 bp away from the transcription start point, with polymorphisms consisting of a change of an A for a G; and the second in gene *ENSG00000157873* (tumor necrosis factor receptor-like 2, *TNFRSF14*), 421 bp within the transcripted region, which corresponds to the first exon of the gene.

The web interface. A web interface to PupaSNP is available at http://pupas.bioinfo.cnio.es/. Lists of genes can be defined by chromosome position, which can be specified in terms of cytoband units or in absolute chromosomal position (as mapped in the corresponding Ensembl assembly). The upstream region makes reference to the number of bases upstream in which TFBSs will be searched for (with a upper limit of 10 000 bp). Also, lists of genes can be uploaded or just pasted into the box. PupaSNP finds all the SNPs mapping to locations that might cause a loss of functionality in the genes. Functional information for the genes can also be obtained from OMIM and from Gene Ontology. Information on homologous genes can also be retrieved. Finally, SNPs do not need to be annotated in the genome to be included in the query tool. The user can specify a list of SNPs using a gene as reference. In this way the use of absolute coordinates, which can easily change between assembly versions, is avoided in favour of the use of coordinates relative to genes, which tend to be more stable. Results include SNPs in a the promoter region of the genes, SNPs located at intron boundaries, SNPs located at exonic splicing enhancers and coding SNPs located at Interpro domains. Figure 1 shows part of the results provided by the program for the SNPs with possible phenotypic effect on genes in the p36.33 cytoband of chromosome 1. Figure 1C is especially interesting because it shows how the scores obtained by the motif scanning method can be used to assess the possible impact of the polymorphism on the recognition of the ESE motif by the cellular machinery.

Both the SNPs and the genes found are linked to the Ensembl Genome Browser.

Experimental validation

The validation status of the SNPs is, in some cases, a much more important factor for their selection than their possible functional role. Such information is scarce: 2 359 534 out of 5 798 183 SNPs in dbSNP build 118 have been validated, which constitutes 40%. However, only 160 466 have estimates of population frequencies and only 94 867 have a phenotype associated. To obtain a sense of the reliability of the SNPs annotated with 'no-info', a set of SNPs was sought for a list of candidate modifier genes related to a phenotype exhibited by *MEN2* (Multiple endocrine neoplasia, type IIA) patients (OMIM, #171400), all of them *RET* mutation carriers. *MEN2* is an autosomal dominant syndrome of multiple endocrine neoplasms, with variable clinical expression even between members of the same family. This fact cannot be explained only by a mutation in a major susceptibility gene, but suggests a role for genetic modifiers, which may also work through quantitative effect.

In most of cases, it was necessary to validate the putative SNPs identified by PupaSNP because there was no information about validation status. To validate SNPs and estimate their allele frequency, 48 non-related individuals from the Spanish population were used. The specific primers used to amplify the fragments of interest by PCR (polymerase chain reaction) were designed using the OLIGO 4.1 program. When possible, the primers were selected and designed to amplify a fragment (200–500 bp) that allowed us to investigate several SNPs at the same time. As a denaturing high-performance liquid chromatograph (dHPLC) system (WAVE, Transgenomics Limited, Crewe, UK) was used for the initial SNP screening, the fragments of interest had a homogeneous GC content across different domains from the DNA fragment to obtain a consistent melting profile. The Navigator software was used for data handling and optimization of the dHPLC system. After normalization, each PCR product that exhibited a change in the chromatogram profile was characterized by sequence analysis. These PCR products were purified using an E.Z.N.A. Cycle-Pure Kit (Omega Bio-tek, USA) according to the manufacturer's instructions, and sequenced using an automatic sequencer ABI PRISM^{TM} 3700 (Applied Biosystems. Perkin Elmer, USA). The reaction was carried out in 4 µl of a Big Dye terminator cycle sequencing Kit (Perkin Elmer, USA), 10 pmol of the sense/antisense primer, 5% DMSO and 6–12 ng of amplified DNA. Although the results obtained here do not pretend to be capable of general extrapolation to the entire database, we have found that 24 out of 28 SNPs assayed proved to be authentic and polymorphic in the Spanish population, which constitutes a good rate.

DISCUSSION

Typically, SNPs have been used as markers to search for the real determinant of a disease in linkage disequilibria with it. As previously mentioned, the use of functional SNPs, which may be the real disease determinants, could be an important factor in increasing the sensitivity of association tests.

Despite the obvious importance that alterations in the regulation, expression level or splicing of genes can have for the phenotype, these have long been ignored in the most common approaches to finding functional SNPs, which have instead focused more on the possible effect of polymorphisms causing amino acid changes. Apart from the databases mentioned above (dbSNP and HGVbase), there are a number of resources

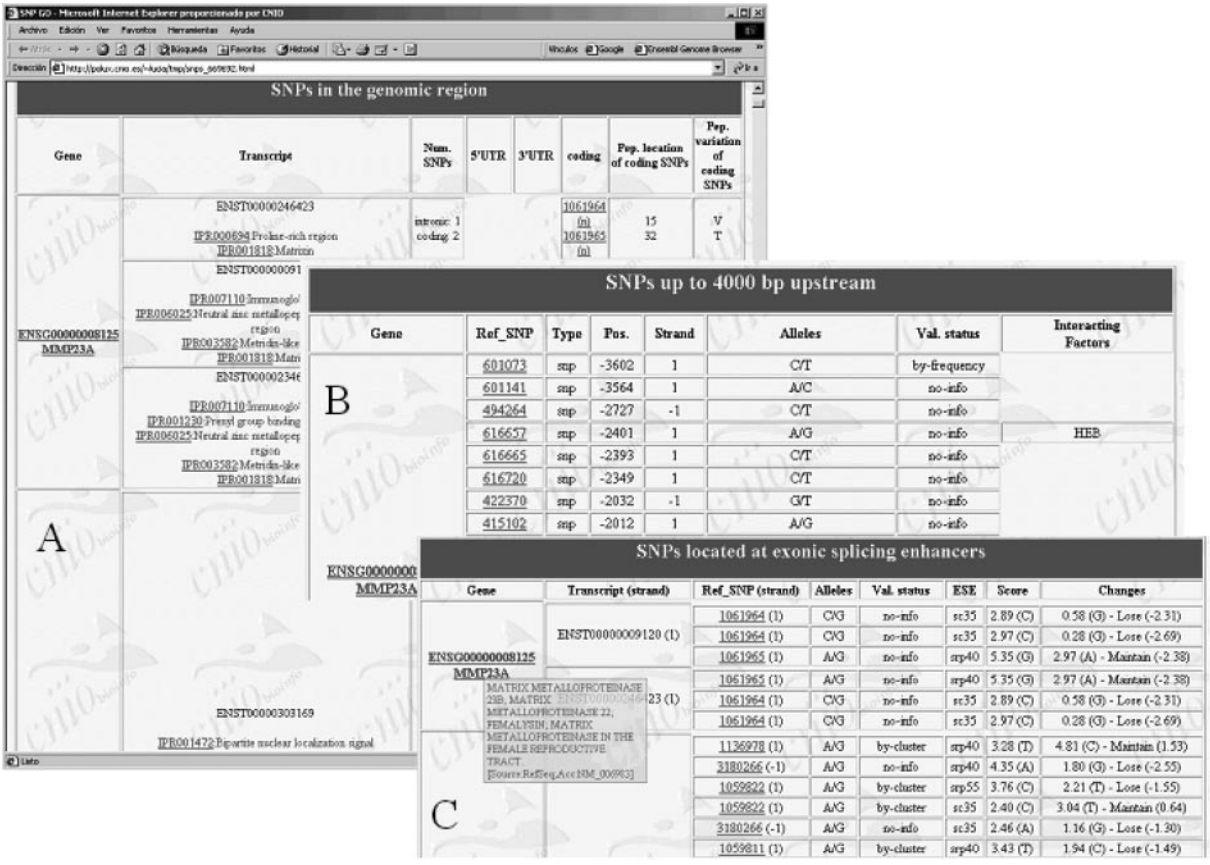


Figure 1. A selection of results from PupaSNP. (**A**) List of genes and the corresponding transcripts with the SNPs mapping to the different regions, which include coding and 5′- and 3′-untranslated regions. For coding SNPs, the position within the transcript and the change produced (if any) is reported. (**B**) SNPs located in the promoter regions (in the example, a limit of 4000 bp was chosen). Disruptions of predicted TFBSs are listed. The validation status of the SNPs ('no-info', 'by-submitter', 'by-frequency', 'by-cluster'; see dbSNP web page) is also provided. (**C**) SNPs located at exonic splice enhancers. The scores make reference to the closeness of the site to the motif. If the polymorphism gives a site with a worst score, this would, generally speaking, probably imply worst recognition of the site by the cellular machinery and, consequently, a putative alteration in the normal splicing process. When the cursor is over the gene name, additional information is displayed.

available over the net collecting information on phenotypes associated with SNPs, such as The Human Gene Mutation Database (http://www.hgmd.org) at the University of Wales, which classifies SNPs according the lesion they cause (missense substitutions, splice variants, and so on) (32) and PicSNP, a catalogue of non-synonymous SNPs obtained from the human genome assembly (33). However, these are mainly specialized catalogues collecting information on SNPs rather than tools for their selection.

PupaSNP constitutes a tool for selecting SNPs with putative phenotypic effects designed for high-throughput experiments. It deals with lists of genes, instead of focusing on individual genes. In addition, more information on different possible motifs with regulatory function has been included. For example, SNPs in ESE had never previously been included in any catalogue.

Multigenic diseases are generally associated with disruptions in proteins that participate in a protein complex or a pathway (4). The inclusion in PupaSNP of information regarding the participation of genes in signalling cascades or in pathways or in protein complexes will be considered in the near future. Databases containing protein interaction data, such as DIP and BIND (see http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html), can be an important source of information to be considered in the search for SNPs affecting multigentic traits.

Despite the fact that PupaSNP is more focused on SNPs with possible effects at transcriptional level, the inclusion of an algorithm for improving the predictions of the effect of SNPs in the proteins, such as FoldX (12), would provide, within the same framework, both types of result.

Minimum SNP set selection allows the user to optimize the number of SNPs required to represent haplotype diversity, thus reducing the cost of genotyping by assaying the minimum number of SNPs required. The inclusion of information on linkage disequilibrium or on haplotype blocks can assist in a more efficient selection of SNPs. Some programs, such as HapScope (34), include information on haplotypes and use them to select minimum subsets of SNPs. Another important issue is the reliability of the SNPs. As previously mentioned, only 40% of the SNPs in dbSNP have been validated, and only for 5% are population frequencies are available. This means that most of the SNPs found in any kind of selection will lack information on their possible presence in the population of interest as a manageable polymorphism. Even though our results suggest a high rate of authenticity, even for the SNPs labeled as 'no-info', they must be treated carefully

and cannot be directly extrapolated to the entire database. As population frequencies are included in the database, these data could be of interest for use as part of the selection process of SNPs

PupaSNP will be the tool used in the first step of the pipeline for the study of polymorphisms at the Spanish National Genotyping Centre (CeGen). For this reason it has been developed to cope with high-throughput experimental designs. PupaSNP takes as input lists of genes (or generates them from chromosomal coordinates) and provides results which integrate all the information available as well as obtained by means of predictions of SNPs with possible functional consequences.

ACKNOWLEDGEMENTS

L.C. and this work are supported by grant PI020919 from the Fondo de Investigaciones Sanitarias. F.A.-S. is supported by grant BIO2001-0068 from Ministerio de Ciencia y Tecnología. This work is also partly supported by a grant from Fundació La Caixa and by the Spanish National Genotyping Centre (CeGen), funded by Genoma España, which is using this program for high-throughput SNP selection.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.*, **33**, 228–237.
4. Badano,J.L. and Katsanis,N. (2002) Human genetics and disease: beyond Mendel: an evolving view of human genetic disease transmission. *Nature Rev. Genet.*, **3**, 779–789.
5. Strittmatter,W.J., Saunders,A.M., Schmechel,D., Pericak-Vance,M., Enghild,J., Salvesen,G.S. and Roses,A.D. (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer's disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.
6. Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M., Binder,V., Finkel,Y., Cortot,A., Modigliani,R., Laurent-Puig,P., Gower-Rousseau,C., Macry,J., Colombel,J.F., Sahbatou,M. and Thomas,G. (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
7. Brookes,A.J., Lehvaslaiho,H., Siegfried,M., Boehm,J.G., Yuan,Y.P., Sarkar,C.M., Bork,P. and Ortigao,F. (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res.*, **28**, 356–360.
8. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W., Kondrashov,A.S. and Bork,P. (2000) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
9. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
10. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.*, **10**, 2319–2328.
11. Chasman,D. and Adams,R.M. (2001) Predicting functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
12. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
13. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.*, **315**, 771–786.
14. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.*, **3**, 285–298.
15. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
16. Montera,M., Piaggio,F., Marchese,C., Gismondi,V., Stella,A., Resta,N., Varesco,L., Guanti,G. and Mareni,C. (2001) A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *J. Med. Genet.*, **38**, 863–867.
17. Yan,H., Yuan, W., Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
18. Prokunina,L., Castillejo-Lopez,C., Oberg,F., Gunnarsson,I., Berg,L., Magnusson,V., Brookes,A.J., Tentler,D., Kristjansdottir,H., Grondal,G., Bolstad,A.I., Svenungsson,E., Lundberg,I., Sturfelt,G., Jonssen,A., Truedsson,L., Lima,G., Alcocer-Varela,J., Jonsson,R., Gyllensten,U.B., Harley,J.B., Alarcon-Segovia,D., Steinsson,K. and Alarcon-Riquelme,M.E. (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. *Nature Genet.*, **32**, 666–669.
19. Harendza,S., Lovett,D.H., Panzer,U., Lukacs,Z., Kuhnl,P. and Stahl,R.A. (2003) Linked common polymorphisms in the gelatinase a promoter are associated with diminished transcriptional response to estrogen and genetic fitness. *J. Biol. Chem.*, **278**, 20490–20499.
20. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.*, **12**, 2249–2254.
21. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nature Genet.*, **33**, 439–440.
22. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
23. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T., Durbin,R., Eyras,E., Gilbert,J., Hammond,M., Hubbard,T., Kasprzyk,A., Keefe,D., Lehvaslaiho,H., Iyer,V., Melsopp,C., Mongin,E., Pettett,R., Potter,S., Rust,A., Schmidt,E., Searle,S., Slater,G., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Stupka,E., Ureta-Vidal,A., Vastrik,I. and Birney,E. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
24. Liu,H.X., Zhang,M. and Krainer, A.R. (1998) Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. *Genes Dev.*, **12**, 1998–2012.
25. Schaal,T.D. and Maniatis,T. (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell Biol.*, **19**, 261–273.
26. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences *Nucleic Acids Res.*, **31**, 3576–3579.
27. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
28. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
29. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P., Copley,R.R., Courcelle,E., Das,U., Durbin,R., Falquet,L., Fleischmann,W., Griffiths-Jones,S., Haft,D., Harte,N., Hulo,N., Kahn,D., Kanapin,A., Krestyaninova,M., Lopez,R., Letunic,I., Lonsdale,D., Silventoinen,V., Orchard,S.E., Pagni,M., Peyruc,D.,

Ponting,C.P., Selengut,J.D., Servant,F., Sigrist,C.J., Vaughan,R. and Zdobnov,E.M. (2003) The InterPro Database brings increased coverage and new features *Nucleic Acids Res*., **31**, 315–318.

30. Al-Shahrour Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
31. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders *Nucleic Acids. Res*., **30**, 52–55.
32. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat*., **21**, 577–581.
33. Chang,H. and Fujita,T. (2001) PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. *Biochem. Biophys. Res. Commun*., **287**, 288–291.
34. Zhang,J., Rowe,W.L., Struewing,J.P. and Buetow,K.H. (2002) HapScope: a software system for automated and visual analysis of functionally annotated haplotypes *Nucleic Acids Res*., **30**, 5213–5221.

Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W501–W505***
doi:10.1093/nar/gki476

PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes

Lucía Conde[1], Juan M. Vaquerizas[1], Carles Ferrer-Costa[2], Xavier de la Cruz[2,4], Modesto Orozco[2,3,5] and Joaquín Dopazo[1,6,*]

[1]Bioinformatics Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid 28029, Spain, [2]Molecular Modelling and Bioinformatics Unit, Institut de Recerca Biomèdica and [3]Structure and Modelling Node INB, Parc Científic de Barcelona, Barcelona 08028, Spain, [4]Institució Catalana per la Recerca i Estudis Avançats (ICREA), 08018 Barcelona, Spain, [5]Departament de Bioquímica i Biología Molecular Facultat de Química, Universitat de Barcelona, Barcelona 08028, Spain and [6]Functional Genomics Node, National Institute of Bioinformatics (INB), CIPF Valencia 46013, Spain

Received February 14, 2005; Revised and Accepted April 15, 2005

ABSTRACT

We have developed a web tool, PupasView, for the selection of single nucleotide polymorphisms (SNPs) with potential phenotypic effect. PupasView constitutes an interactive environment in which functional information and population frequency data can be used as sequential filters over linkage disequilibrium parameters to obtain a final list of SNPs optimal for genotyping purposes. PupasView is the first resource that integrates phenotypic effects caused by SNPs at both the translational and the transcriptional level. PupasView retrieves SNPs that could affect conserved regions that the cellular machinery uses for the correct processing of genes (intron/exon boundaries or exonic splicing enhancers), predicted transcription factor binding sites and changes in amino acids in the proteins for which a putative pathological effect is calculated. The program uses the mapping of SNPs in the genome provided by Ensembl. PupasView will be of much help in studies of multifactorial disorders, where the use of functional SNPs will increase the sensitivity of the identification of the genes responsible for the disease. The PupasView web interface is accessible through http://pupasview.ochoa.fib.es and through http://www.pupasnp.org.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and, with the recent availability of high-throughput methodologies, are considered one of the most powerful tools in the search for e.g. disease susceptibility genes and drug response-determining genes (1,2). However, complex diseases, for which markers display weak associations, still constitute a challenge. Most probably, advancement in the knowledge of such diseases will come from improved genotyping methods in combination with the proper bioinformatics design strategies (3).

It is generally believed that multigenicity reflects disruptions in proteins that participate in a protein complex or in a pathway (4). Typically, SNPs have been used as markers; that is, the real determinant of the disease was not the SNP itself but some other mutation in linkage disequilibrium (LD) with it. Because of this, the use of functional SNPs could be an important factor in increasing significantly the sensitivity of association tests. In fact, several complex genetic disorders such as Alzheimer's disease (5) and Crohn' disease (6) have been associated with functional SNPs, lending weight to strategies giving priority to candidate markers based upon predictable function. Several estimations suggest that, on average, some 20% of SNPs could directly damage proteins (7).

Much attention has been focused on modelling by different methods the possible phenotypic effect of SNPs that cause

*To whom correspondence should be addressed. Email: jdopazo@ochoa.fib.es
Present address:
Lucia Conde, Juan M. Vaquerizas and Joaquín Dopazo, Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Valencia 46013, Spain

amino acid changes (7–13), and only recently has interest focused on functional SNPs affecting regulatory regions or the splicing process (14). However, there is increasing evidence that many human disease genes are the result of exonic or non-coding mutations affecting regulatory regions (15–17). A recent large-scale screening over a set of 16 chromosomes found SNPs in the promoter regions of 35% of the genes, and experimental evidence suggested that around a third of promoter variants may alter gene expression to a functionally relevant extent (18). Alternative splicing produced by mutations in intron/exon junctions, or in distinct binding motifs, such as exonic splicing enhancers (ESEs) (19), has also been related to different diseases (20). In fact, it has been estimated that 15% of point mutations that result in human genetic diseases cause RNA splicing defects (21).

In addition to functional information, population frequency is another important factor to be taken into account when selecting SNPs. Thus, infrequent polymorphisms will be of scarce interest as markers. Also, LD is another interesting factor in selecting SNPs as markers since, if two SNPs are in strong LD, only one of them will provide enough information for any association or linkage test.

With the idea of selecting optimal sets of SNPs using as much information as possible on putative phenotypic effect, population frequencies and LD, we have developed PupasView (Putative Phenotypic Alterations caused by SNPs Viewer), a server that can be used alone or in combination with PupaSNP (14).

PupasView works not only as a viewer of where SNPs are located, but also as a selector in which different filters based on combinations of functionality and population frequencies can be interactively applied over the LD parameters in order to obtain an optimal selection of SNPs for genotyping studies, in such a way that with a minimum number of SNPs maximum information on the genic region is obtained.

Criteria to consider an SNP a good candidate for genotyping studies

There are three important properties for an SNP to be considered an optimal candidate for genotyping purposes: functional effect, minor allele frequency and LD with respect to other SNPs. Finding such optimal SNPs is not always possible, but the idea behind PupasView is to facilitate the selection process in order to achieve a final collection of SNPs bearing the maximum amount of information. PupasView works as an SNP selector. Different filters can be interactively applied to the LD information available based on distinct functional properties, cross-species conservation and population frequency. This permits a final selection of a minimum number of SNPs with optimal properties in terms of population frequencies and potential phenotypic effect.

Finding SNPs with potential phenotypic effect

PupasView uses a precompiled database which contains a collection of dbSNP entries mapped to the Golden Path genome assembly, as implemented in the human section of Ensembl (http://www.ensembl.org). Part of this database is common to the PupaSNP program (14). The SNPs have been labelled according to their potential effects on the phenotype. We have taken into account both transcriptional and gene product levels. Regions 10 000 bp upstream of the genes belonging to the promoter region of each gene in the list have been scanned for the presence of possible different regulatory motifs. These include alterations in:

(i) *Transcription factor binding sites*. Promoter regions were scanned for the presence of possible transcription factor binding sites. The program Match (22) was used for this purpose, using only high-quality matrices and with a cut-off to minimize false positives from the Transfac database (23). SNPs located within these motifs are considered to have a putative phenotypic effect in the expression of the gene. Almost four million such motifs were found, with 130 373 SNPs mapping onto them.

(ii) *Intron/exon border consensus sequences*. Ensembl APIs (24) were used to extract the intron/exon organization of the genes and the corresponding sequences. The two conserved nucleotides on each side of the splicing point, which constitute the splicing signal (21), were then located and all the SNPs altering these signals were recorded. More than 700 000 intron/exon boundaries could be defined in human genes with 1786 SNPs mapping onto them.

(iii) *ESEs*. Mutations that inactivate or activate an ESE sequence may result in exon skipping, errors in alternative splicing patterns, malformation and so on. Different classes of ESE consensus motifs have been described, but they are not always easily identified. Exon sequences were scanned to identify putative ESEs responsive to the human SR proteins SF2/ASF, SC35, SRp40 and SRp55, using the available weight matrices (20). A score was obtained that is related to the likelihood that the site found is a real ESE. Only ESE sites with scores over the threshold [see (20) for details] were taken into account in the analysis. More than 11 million ESEs were found, with 299 106 SNPs located in them.

(iv) *Triplex-forming oligonucleotide target sequences* (TTSs). It has been found that the population of TTSs is much more numerous than expected from simple random models (25). The population of TTSs is large in the whole genome, without major differences between chromosomes, but with a large concentration in regulatory regions, especially in promoter zones, which suggests a tremendous potential for triplex strategy in the control of gene expression (25). Although the role of TTSs in regulation is still a matter of speculation, the program also reports SNPs disrupting these structures. Some 5.4 million putative triplex-forming sequences were found, and 364 314 SNPs mapped onto them.

(v) *SNPs in exons that cause an amino acid change*. Any SNP causing a change of amino acid, independent of any speculation on its possible phenotypic effect, is reported. There are 45 906 such SNPs.

(vi) *SNPs in exons that cause an amino acid change with putative pathological effect*. The putative pathological effect of an amino acid change can be predicted using neural networks (NNs) carefully trained to predict disease-associated amino acidic polymorphism (12,13). The server implements a small NN (1 hidden layer and 20 nodes) and three sequence-derived descriptors (PAM40, PSSM and variability), which are either retrieved from databases or determined internally from multiple alignments using

two-iterations PSI-Blast (26) run over a non-redundant SwissProt/TrEMBL database. The trained method displays a success rate >80% in cross-validation experiments. According to the algorithm, 19 309 SNPs displayed a high probability of having pathological effect.

(vii) *Human–mouse conserved regions*. Untranslated whole genome comparisons by BLASTZ were performed for species pairs which are thought to be similar enough to be able to detect homology directly at the DNA level (27). Of particular interest is mouse (or rat) because of its phylogenetic position with respect to humans: distant enough to interpret conservation as important but not so distant as to lose most of the similarity. The phenotypic effect of a change in such regions is quite speculative, but cross-species conservation can be useful in cases in which no other information is available. It is also useful for reinforcing the likelihood of other predictions (e.g. an ESE in a conserved region is more likely to be real than one in a non-conserved region).

Frequency information and validation status

There are >10 million SNPs stored in the last build of dbSNP (build 124), and more than half of these have been validated by different means (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). Validation status is annotated and is an important field in terms of trusting an SNP. But, in addition to being real, an SNP must exist in the population at frequencies which make it a suitable marker. Very infrequent SNPs are not suitable for association or linkage studies. For almost half a million SNPs frequency data in different populations are available.

Blocks and LD parameters

LD measures the correlation between two neighbouring genetic variants in a specific population. The program HaploView (28) is used to infer blocks using different procedures. In one of the most common procedures (29), 95% confidence bounds based on the D' LD parameter are generated and each comparison is called 'strong LD', 'inconclusive' or 'strong recombination'. A block is created if 95% of informative (i.e. non-inconclusive) comparisons are 'strong LD'. A block can be considered a region with a low recombination rate. Ideally, a block could properly be described by a unique SNP. Two other methods are used: the four gamete rule (30) and the Solid Spine of LD (28). Blocks are displayed in the bottom of the PupasView window. Also D', R^2 and LOD parameters between adjacent SNPs can be visualized by placing the cursor between them. Only HapMap genotyped SNPs (31) are used to calculate blocks and LD parameters.

The web interface of the SNPs selector

The main purpose of PupasView is to provide the user with an optimal set of SNPs for genotyping experiments by filtering the annotated SNPs using a series of filters related to their impact in protein functionality and pathology, their population frequency and LD.

The input is a gene identifier (Ensembl IDs or external IDs, which include GenBank, Swissprot/TrEMBL and other gene IDs supported by Ensembl). The program can also be invoked from PupaSNP. The program presents a list of options that can be selected and applied as many times as desired. The options include

- Validation status obtained from dbSNP
- Type of SNP (coding, intron, untranslated region, local), according to its position in the gene
- Frequency and population, an option that allows the possibility of filtering by a range of frequencies of the minor allele in one or more populations (Europe; Europe, multinational; Europe, North America; North America; Central/South America; North/East Africa and Middle East; Central/South Africa; West Africa; Central Asia; East Asia; Pacific; multinational; unknown; HapMap)
- Functional properties as follows:
  - non-synonymous SNPs [all or only those predicted as pathological by the pmut algorithm (12,13)]
  - SNPs disrupting predicted transcription factor binding sites (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting predicted ESEs (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting potential triplex-forming regions (all or only those that are in regions conserved in the mouse genome)
  - SNPs disrupting intron/exon boundaries
  - regions conserved in mouse
- Options for the way in which blocks are constructed:
  - confidence intervals (29)
  - four gamete rule (30)
  - Solid Spine of LD (28).

Figure 1 shows the view of the results. The viewer of PupasView has been constructed using Ensembl APIs (24). Figure 1A shows the result of running PupasView on the gene TP53 without applying any filter. All the SNPs in the gene and the neighbourhood are displayed. If the cursor is over an SNP, information on it is displayed by means of pop-up text. Figure 1B shows a subselection of these SNPs obtained after selecting only SNPs for which population frequency was available. Finally, Figure 1C shows the selection obtained if only SNPs with putative functional effect are chosen. This will constitute the final, reduced subset of optimal SNPs. The upper horizontal bar below the figure represents LD parameters (which can be individually obtained by placing the cursor over them). The lower horizontal bar represents the block found with the selected algorithm. The blocks are displayed graphically with brown rectangles going from the first to the last SNP within the block. When the cursor is over the rectangles, a tooltip text pops up in the block showing the SNPs and the haplotypes (with HapMap frequencies in parentheses). Tag SNPs are signalled with an exclamation mark (!).

DISCUSSION

It is believed that improved genotyping methods in combination with the proper bioinformatics design strategies will offer better opportunities for the study of complex diseases (3). The use of functional SNPs could be an important factor in increasing the sensitivity of association tests. Different bioinformatics approaches have been focused mainly on the effect of coding SNPs, but also recently on SNPs affecting the regulation or the splicing of genes (14).

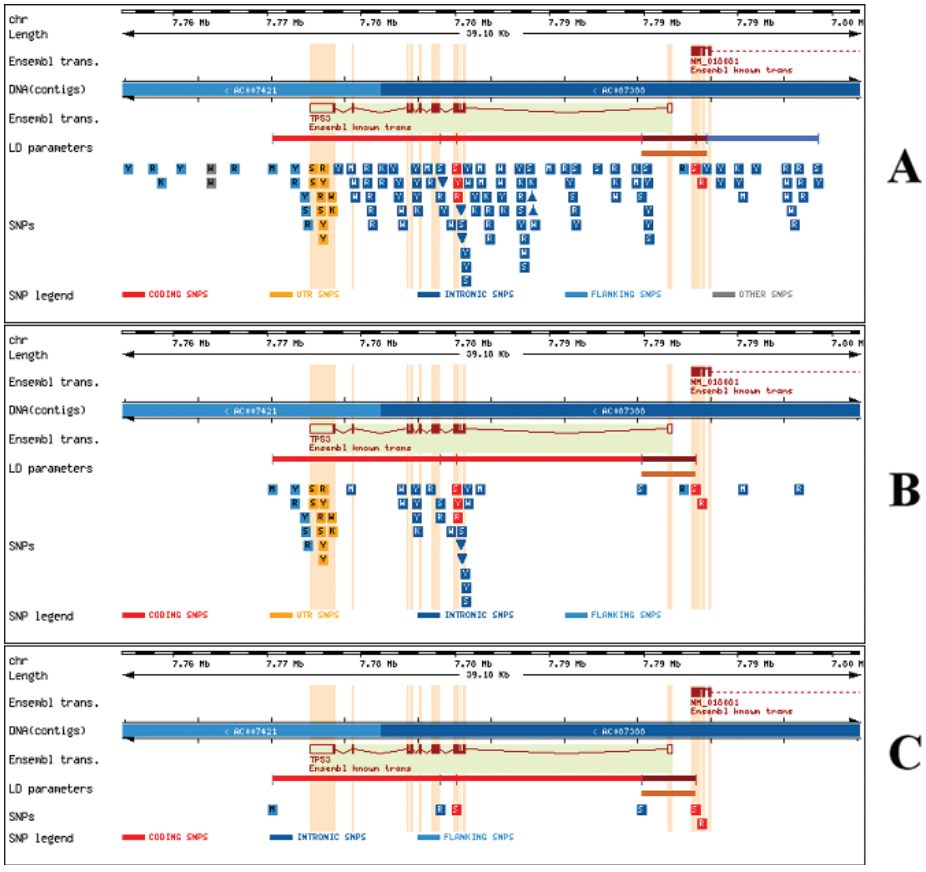


Figure 1. Sequential application of filters in PupasView. (**A**) SNPs in gene TP53. (**B**) SNPs together with population frequencies. (**C**) SNPs with any functional characteristic. Depending on the versions of Ensembl and dbSNP, the appearance of the figure can change.

PupasView is the first tool that integrates both transcriptional and translational phenotypic effects caused by polymorphisms. It provides an interactive environment in which functional information and population frequency data can be used over LD parameters as sequential filters to obtain a final list of SNPs optimal for genotyping purposes.

PupasView is closely linked to our previous program PupaSNP (14), which is a tool for selecting SNPs with putative phenotypic effects. PupaSNP, designed for high-throughput experiments, has been used to design >9000 sets of SNPs, and has a daily average of 50 uses. PupasView assists in the last refinement step of gene-by-gene selection of SNPs. Figure 1 illustrates the effect of applying successive filter steps, which are, conceptually, first to select only those SNPs which are real (with reported population frequencies) and then to select only functional SNPs. In the last view (Figure 1C), LD parameters can be used to help in the final selection.

More than 5000 SNPs have been selected using PupaSNP and PupasView in the first step of the pipeline for the study of polymorphisms at the Spanish National Genotyping Centre (CeGen).

ACKNOWLEDGEMENTS

L.C. and this work are supported by grant PI020919 from the FIS. J.M.V. is supported by the FPU fellowship programme from the MEC. This work is also partly supported by a grant from the Fundació La Caixa and the Fundación Ramón Areces. The Functional Genomics and Structure and Modelling nodes of the INB are funded by the Fundación Genoma España. CeGen, also funded by the Fundación Genoma España, is currently using the PupaSNP and PupasView programs for high-throughput SNP selection. Funding to pay the Open Access publication charges for this article was provided by Fundación Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.*, **33**, 228–237.
4. Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.*, **3**, 779–789.
5. Strittmatter,W.J., Saunders,A.M., Schmechel,D., Pericak-Vance,M., Enghild,J., Salvesen,G.S. and Roses,A.D. (1993) Apolipoprotein E: high-vidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.

6. Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
7. Sunyaev,S., Ramensky,V., Koch,I., Lathe,W., Kondrashov,A.S. and Bork,P. (2000) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
8. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
9. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.*, **10**, 2319–2328.
10. Chasman,D. and Adams,R.M. (2001) Predicting functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
11. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
12. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.*, **315**, 771–786.
13. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.
14. Conde,L., Vaquerizas,J.M., Santoyo,J., Al-Shahrour,F., Ruiz-Llorente,S., Robledo,M. and Dopazo,J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**, W242–W248.
15. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
16. Yan,H., Yuan,W., Velculescu,V.E., Vogelstein,B. and Kinzler,K.W. (2002) Allelic variation in human gene expression. *Science*, **297**, 1143.
17. Prokunina,L., Castillejo-Lopez,C., Oberg,F., Gunnarsson,I., Berg,L., Magnusson,V., Brookes,A.J., Tentler,D., Kristjansdottir,H., Grondal,G. *et al.* (2002) A regulatory polymorphism in *PDCD1* is associated with susceptibility to systemic lupus erythematosus in humans. *Nat. Genet.*, **32**, 666–669.
18. Hoogendoorn,B., Coleman,S.L., Guy,C.A., Smith,K., Bowen,T., Buckland,P.R. and O'Donovan,M.C. (2003) Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.*, **12**, 2249–2254.
19. Colapietro,P., Gervasini,C., Natacci,F., Rossi,L., Riva,P. and Larizza,L. (2003) NF1 exon 7 skipping and sequence alterations in exonic splice enhancers (ESEs) in a neurofibromatosis 1 patient. *Hum. Genet.*, **113**, 551–554.
20. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat. Rev. Genet.*, **3**, 285–298.
21. Krawczak,M., Reiss,J. and Cooper,D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.*, **90**, 41–54.
22. Kel,A.E., Gössling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
23. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
24. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
25. Goni,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.*, **32**, 354–360.
26. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
27. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ [Erratum (2004) *Genome Res.*, **14**, 786.]. *Genome Res.*, **13**, 103–107.
28. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
29. Gabriel,S.B., Schaffner,S.F., Nguyen,H, Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2259.
30. Wang,N., Akey,J.M., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination and mutation. *Am. J. Hum. Gen.*, **71**, 1227–1234.
31. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W621–W625***
doi:10.1093/nar/gkl071

PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes

Lucía Conde[1], Juan M. Vaquerizas[1], Hernán Dopazo[1], Leonardo Arbiza[1], Joke Reumers[2], Frederic Rousseau[2], Joost Schymkowitz[2] and Joaquín Dopazo[1,3,*]

[1]Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46013, Spain, [2]Switch laboratory, Flanders Interuniversity Institute for Biotechnology. (VIB), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium and [3]Functional Genomics Node, INB, CIPF Valencia 46013, Spain

Received February 14, 2006; Revised February 23, 2006; Accepted March 3, 2006

ABSTRACT

We have developed a web tool, PupaSuite, for the selection of single nucleotide polymorphisms (SNPs) with potential phenotypic effect, specifically oriented to help in the design of large-scale genotyping projects. PupaSuite uses a collection of data on SNPs from heterogeneous sources and a large number of pre-calculated predictions to offer a flexible and intuitive interface for selecting an optimal set of SNPs. It improves the functionality of PupaSNP and PupasView programs and implements new facilities such as the analysis of user's data to derive haplotypes with functional information. A new estimator of putative effect of polymorphisms has been included that uses evolutionary information. Also SNPeffect database predictions have been included. The PupaSuite web interface is accessible through http://pupasuite.bioinfo.cipf.es and through http://www.pupasnp.org.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and constitute one of the most powerful tools in the search for disease susceptibility genes, drug response-determining genes and the like (1,2). With the introduction of large-scale genotyping techniques the bottleneck in this type of experiments has moved towards the management and analysis of the data generated. In this context, one of the topics which has become a problem is the step of the selection of the optimal set of SNPs (among several thousands of candidates in some cases) for the genotyping experiment. Optimal SNPs must be the best possible markers for traits, which often are multigenic, usually reflecting disruptions in proteins that participate in a protein complex or a in a pathway (3). Unfortunately, complex multigenic traits, for which markers display weak associations, still constitute a challenge. Factors such as linkage disequilibrium (LD) and minor allele frequency (MAF) are of major importance for selecting optimal candidate SNPs. Recently, the predicted functional effect of an SNP is gaining importance as a selection criterium because it constitutes a potential important factor for increasing the sensitivity of association tests significantly (3–6). The availability of information on LD from projects such as HapMap (7), on MAFs (8) and improved methods for predicting function (5,6,9), allow for a more sophisticated selection of candidate SNPs beyond the classical one-SNP-at-a-time approach. Thus, SNPs can be selected taking into account the evolutionary constraints of the region analysed along with its likelihood of being the causative agent of any type of damage. Algorithms which use information to facilitate the posterior analysis of the results, such as the estimation of haplotype blocks (10), combined with functional prediction of the effect of the SNPs, are expected to have a major impact on the efficiency of a large-scale genotyping study. PupaSuite belongs to this new generation of tools. PupaSuite combines the facilities offered by PupaSNP (6) and PupasView (5) with new algorithms and visualisation procedures for functional haplotype prediction. The PupaSNP and PupasView programs are part of the pipeline of genotyping of the Spanish National Genotyping Center (CeGen; http://www.cegen.org/). Both tools combined bear an average of 60 SNP designs per day.

OUTLINE OF THE PROGRAM

PupaSuite combines the functionality of PupaSNP (6) and PupasView (5) in a unique and more integrated interface, and adds new modules to facilitate the selection of the optimal set of SNPs for a large-scale genotyping study. Following the philosophy of PupaSNP, the program allows to input either *lists of genes* or *chromosomal regions*, which would

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

correspond to two common types of analysis: genes probably related to a disease because they are functionally related (e.g. they belong to a pathway affected in the disease), or genes present in a chromosomal region linked to a disease. PupaSuite can also directly analyse *lists of SNPs*. In these three cases a list of SNPs with their putative functional effect is reported. In the case of chromosomal regions it is also possible to find haplotype blocks (10). For the list of SNPs, in addition to their putative functional effect, it is possible to retrieve information on MAF in different populations from dbSNP (8) [as annotated in the Ensembl (11)], as well as LD parameters and haplotype blocks.

In addition to the analysis of lists of SNPs there is another new option: *Functional haplotypes*. This option (see below) allows the user to test their own SNP data and to find haplotypes (12) with the functional SNPs (5,6) and the tag SNPs (13) highlighted. Case-control studies can also be performed at this stage. The option *Display and Filter SNPs for a single gene* implements new functionalities in an environment *a la* PupasView (5). More information is presented in a graphical intuitive format (Figure 1). This option allows the sequential and interactive application of filters based on functionality, conservation, MAF and the like (5) thus permitting an easy selection of a set of optimal SNPs for a particular gene.

CRITERIA TO SELECT SNPS AS A GOOD CANDIDATES FOR GENOTYPING

Here three important features of a SNP have been taken into account in order to be considered as an optimal candidate for genotyping purposes: MAF, LD with respect to other candidates (5) and putative functional effect. MAF values were taken from the Ensembl (11), which maps dbSNP (8) data onto the corresponding chromosomal coordinates. LD are calculated as r^2 and D′ with the Haploview program (14). The putative functional effect has been estimated in both coding and non-coding regions as described in (5). The following features have been used to report the putative functional effect of a polymorphism in non-coding nucleotides:

(i) Transcription factor binding sites from the Transfac database (15).
(ii) Intron/exon border consensus sequences.
(iii) Exonic splicing enhancers (16).
(iv) Triplex-forming oligonucleotide target sequences (17).

Regarding the putative impact of a cSNP, the following data and estimators are reported:

(i) SNPs in exons causing an amino acid change (purely a list of cSNPs)
(ii) Pmut (18,19) predictions.
(iii) Selective strengths (ω parameter). This estimator is new in this version of the program (see below)
(iv) SNPeffect (9,20,21) predictions. New in this version of the program (see below).

The likelihood of the predictions can be reinforced by looking simultaneously for human-mouse conserved regions (22) as reported in Ensembl (http://www.ensembl.org).

EVOLUTION AT WORK: THE SELECTIVE STRENGTHS ON CSNPS

The combined effect of all the selective pressures causes the preservation of the functionally relevant parts of the genes. Under this perspective, comparative and evolutionary studies have been used to predict the putative functional effect of SNPs (19,23) although these have mainly ignored the underlying phylogeny. Here we present another more accurate estimator of functional effect, based on sequence comparison, but taking into account phylogenetic information (24). The selective pressures acting at a codon-level where non-synonymous cSNPs are found were evaluated by means of two alternative approaches: codon-based maximum likelihood (ML) models (25) implemented in PAML (26), and likelihood-ratio (SLR) method (27) for testing deviations of neutrality.

Under the first approximation, an a priori statistical distribution describing the variation of $\omega = dN/dS$ among sites is assumed for a number k of different classes of sites with ω_k values at a proportion p_k of the sequences representing the effects of purifying selection ($0 < \omega_0 < 1$), neutral evolution ($\omega_1 = 1$), and positive selection ($\omega_2 > 1$) (25). The method involves two main steps: first, the adjustment by maximum likelihood of the evolutionary parameters to the sequences of the species compared considering two different models; and second, the use of the Bayes theorem to compute the posterior probability that each site belongs to a specific site class ω_k defined under an a priori distribution (28). Two different models (M2a and M8) were evaluated by maximum likelihood on the sequences (29).

Under the sitewise likelihood-ratio method (SLR) a site-by-site approach to test for neutrality is used. In contrast to similar approaches developed previously (30), SLR uses the entire alignment of the sequence to determine parameters common to all sites, such as evolutionary distances. Using this approach there is no need to specify a model of how ω varies along the sequence. A correction for multiple testing in order to obtain statistical confidence for inferences on deviations from neutrality on each site is also performed.

SNPEFFECT DATABASE

The SNPeffect database (9) describes the effect of coding non-synonymous SNPs on several phenotypic properties of human proteins using either sequence-based or structural bioinformatics tools. Molecular phenotypes are grouped in three categories: structure and dynamics, functional sites and cellular processing. Next to various external tools SNPeffect uses algorithms developed at the collaborating research groups, among which Tango (20) to predict β-aggregation regions in protein sequences and FoldX (21) to predict the stability change caused by the single amino acid variation.

FUNCTIONAL HAPLOTYPES

In addition to using already available data, the users can input their own data to use the predictions on possible functional effects in combination with haplotype analysis. This possibility can be used through the *Functional haplotypes*

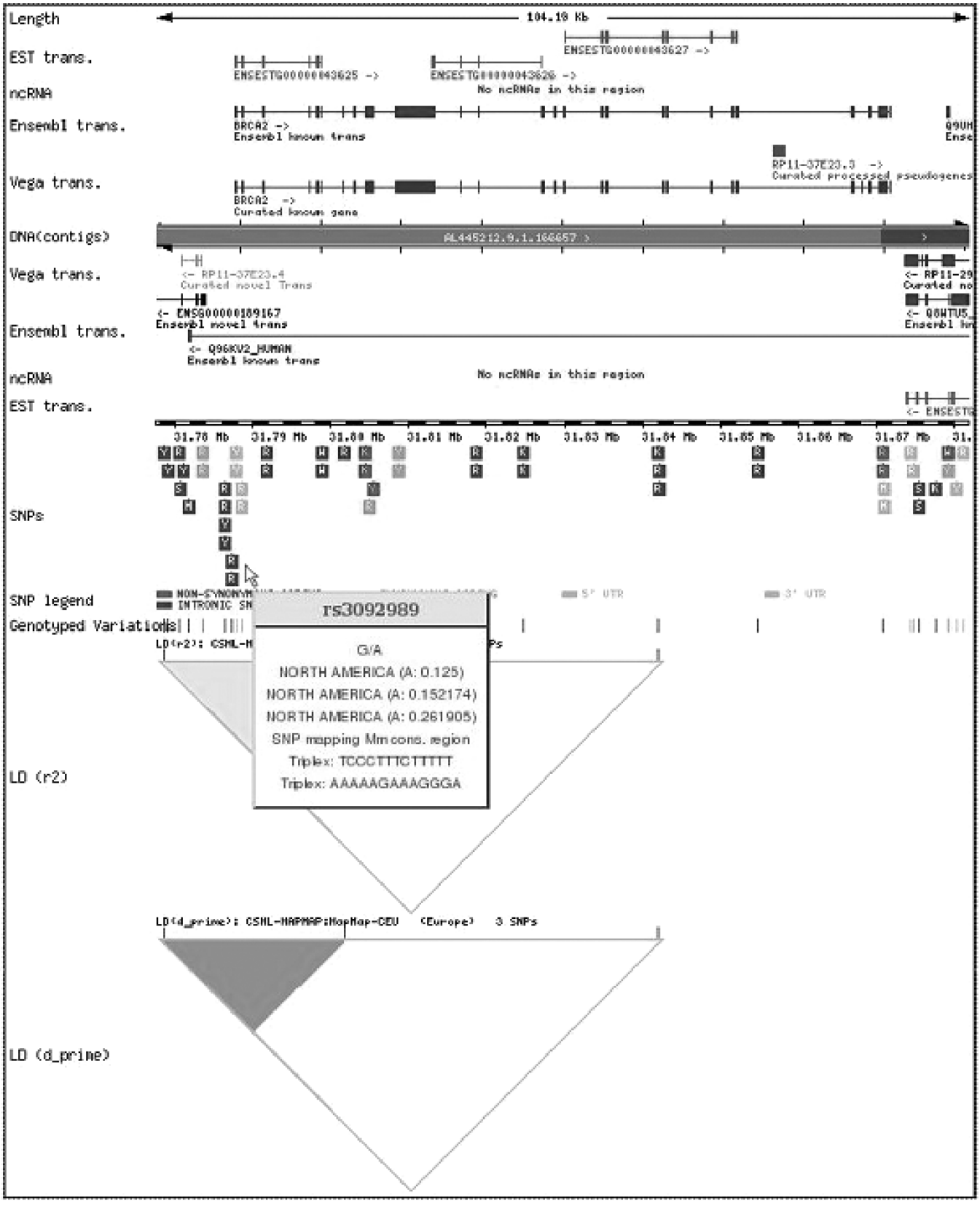


Figure 1. Output with the graphic representation of SNPs with putative functional effect in the gene BRCA2, along with LD maps.

option. Data must be provided to the program in linkage pedigree format (pre MAKEPRED, http://pupasuite.bioinfo.cipf.es/html/help/index.html). The PupaSuite estimates blocks by three methods: Confidence intervals (10), Four gamete rule (31) and Solid Spine of LD (14) and reconstruct haplotypes using the EM algorithm (12) as implemented in Haploview (14). The haplotypes found in this way are represented with the corresponding functional information on all the SNPs included in it and all the LD values. This representation provides a very intuitive picture of the possible functional impact of any of the haplotypes beyond the individual effect of each SNP. For case/control data a chi-square test is performed and the corresponding *P*-value for the allele frequencies in cases versus control is reported. The combination of functional haplotype information with case/control tests allows to easily ascribe cases to haplotypes with functional alterations.

DISCUSSION

We have presented an integrated resource for helping in the selection of optimal sets of SNPs oriented to large-scale genotyping assays. The program merges the functionalities of other two previous resources, PupaSNP (6) and PupasView (5), and expand the capabilities of the program with new information and new facilities. The SNPeffect database (9) as well as a new, unpublished prediction method has been included to improve the estimation of the putative pathological effect of SNPs. Moreover, in addition to use publicly available data on SNPs, users can analyse their own experiments. What is novel and unique to tools of this type is the possibility of analysing functionally haplotypes, beyond the classical analysis one-SNP-at-a-time which ignores interactions between the mutations.

The usefulness of this type of resources is proven by the use made by the CeGen in its pipeline of genotyping. The previous tools, which have been running for more than two years, have now an approximate average of 60 daily SNP designs (http://bioinfo.cipf.es/webalizer/pupasnp and http://bioinfo.cipf.es/webalizer/pupasview).

ACKNOWLEDGEMENTS

This work is supported by grants from Fundació La Caixa, Fundación BBVA, MEC BIO2005-01078 and NRC Canada-SEPOCT Spain. The Functional Genomics node (INB) is supported by Genoma España. LC is supported by fellowship from the CeGen (Genoma España). Funding to pay the Open Access publication charges for this article was provided by Genome España.

Conflict of interest statement. None declared.

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
3. Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. *Nature Rev. Genet*., **3**, 779–789.
4. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet*., **33**, 228–237.
5. Conde,L., Vaquerizas,J.M., Ferrer-Costa,C., de la Cruz,X., Orozco,M. and Dopazo,J. (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res*., **33**, W501–W505.
6. Conde,L., Vaquerizas,J.M., Santoyo,J., Al-Shahrour,F., Ruiz-Llorente,S., Robledo,M. and Dopazo,J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res*., **32**, W242–W248.
7. Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J. and Donnelly,P. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, D173–D180.
9. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J., Stricher,F., Serrano,L. and Rousseau,F. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res*., **33**, D527–D532.
10. Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
11. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
12. Qin,Z.S., Niu,T. and Liu,J.S. (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet*., **71**, 1242–1247.
13. de Bakker,P.I., Yelensky,R., Pe'er,I., Gabriel,S.B., Daly,M.J. and Altshuler,D. (2005) Efficiency and power in genetic association studies. *Nature Genet*., **37**, 1217–1223.
14. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
15. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
16. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet*., **3**, 285–298.
17. Goni,J.R., de la Cruz,X. and Orozco,M. (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res*., **32**, 354–360.
18. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol*., **315**, 771–786.
19. Ferrer-Costa,C., Orozco,M. and de la Cruz,X. (2004) Sequence-based prediction of pathological mutations. *Proteins*, **57**, 811–819.
20. Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol*., **22**, 1302–1306.
21. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res*., **33**, W382–W388.
22. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res*., **13**, 103–107.
23. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet*., **10**, 2319–2328.
24. Arbiza,L., Duchi,S., Montaner,D., Burguet,J., Pantoja-Uceda,D., Pineda-Lucena,A., Dopazo,J. and Dopazo,H. (2006) Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J. Mol. Biol*., in press.

25. Yang,Z. and Nielsen,R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.
26. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
27. Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.
28. Yang,Z., Wong,W.S. and Nielsen,R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.
29. Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**, 431–449.
30. Suzuki,Y. and Gojobori,T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.*, **16**, 1315–1328.
31. Wang,N., Akey,J.M., Zhang,K., Chakraborty,R. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J Hum. Genet.*, **71**, 1227–1234.

Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W81–W85***
doi:10.1093/nar/gkm257

ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling

Lucía Conde[1], David Montaner[1,2], Jordi Burguet-Castell[1], Joaquín Tárraga[1,2], Ignacio Medina[1], Fátima Al-Shahrour[1] and Joaquín Dopazo[1,2,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF) and [2]Functional Genomics Node, INB, CIPF, Valencia 46013, Spain

Received January 30, 2007; Revised March 28, 2007; Accepted April 8, 2007

ABSTRACT

We present the ISACGH, a web-based system that allows for the combination of genomic data with gene expression values and provides different options for functional profiling of the regions found. Several visualization options offer a convenient representation of the results. Different efficient methods for accurate estimation of genomic copy number from array-CGH hybridization data have been included in the program. Moreover, the connection to the gene expression analysis package GEPAS allows the use of different facilities for data pre-processing and analysis. A DAS server allows exporting the results to the Ensembl viewer where contextual genomic information can be obtained. The program is freely available at: http://isacgh.bioinfo.cipf.es or within http://www.gepas.org.

INTRODUCTION

Genetic aberrations, such as losses (deletions) or gains (amplifications) of genetic material that affect certain regions of the genome, have been shown to be on the basis of many human pathologies, including rare diseases, as mental retardation (1), or much more prevalent pathologies, as cancer (2).

Classical approaches to characterize these genetic aberrations used comparative genomic hybridization (CGH), in which genomic DNA was hybridized to metaphase chromosomes (3). Recently, however, the use of different types of microarrays to directly study genomic variations in DNA copy number is becoming more and more popular. Such massive genomic approaches are known as array comparative genomic hybridization, or Array CGH (4). Different options are used to implement Array CGHs including large genomic clones (5), cDNAs (6), oligonucleotides (7) and even SNP genotyping platforms (8). These new technologies along with the use of expression arrays offer for the first time the opportunity of characterize in an accurate way the dependence of gene expression on alterations in genomic copy number (9,10).

As in other high-throughput methodologies, data analysis and, in particular, biological interpretation of the results constitutes a well-known bottleneck. Specific problems related to the analysis of Array CGH can be circumscribed mainly to: (i) the accurate definition of the borders of the genetic alteration and the copy number estimation, (ii) the appropriate mapping and visualization of the data onto the chromosomes and (iii) the possibility of formulating reasonable hypothesis that link genes to diseases by understanding the alteration of the functions at molecular level. The first aspect has been the motivation for a number of analytical approaches recently proposed (11,12). Although several programs have been developed for array-CGH data visualization and analysis, almost all of them are stand-alone applications in different programming languages such as R and MATLAB scripts, C or java (12). To our knowledge only two web-based applications for array-CGH data analysis have been published to date: CAPweb (13) and ArrayCyGHt (14). Among the specific problems previously mentioned, probably, the last one is the most relevant given that the ultimate aim of studies of copy number chromosomal alterations is to understand what is the functional effect produced at molecular level that can help to interpret the pathologic phenotype. In the classical vision, one or a few key genes are the causative factors for this type of pathologies, and the problem consisted in identifying such genes within the region amplified or deleted. This vision is changing by the recent report of regions in the chromosomes of higher eukaryotes containing coexpressing genes (15) which, in addition, are functionally related (16). Actually, regional arrangements of genes have found to be regulated

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

not only by copy number alterations but also by different mechanisms such as epigenetic modifications (17). This reinforces the functional role of chromosomal regions containing groups of functionally related genes and their possible impact on diseases such as cancer (18). This important aspect, however, remains mostly overlooked in the tools for the analysis of copy number alterations.

We present here the ISACGH program that allows visualizing array CGH data or/and expression arrays onto human or mouse chromosomal coordinates (automatically found through their standard identifiers) and represents the regions with copy number alterations found by using different methods. Correlations between copy number and gene expression level can be visualized in different plots. The program finds minimal common regions with altered copy number across different arrays. Although ISACGH can be used alone, it is tightly integrated into the GEPAS (19,20) and Babelomics (21) packages. Thus, normalization and any other data transformation operations can directly be performed within a common environment, without the necessity of reformatting the data. The connection of ISACGH to different tools for functional profiling (21,22) offer the possibility of studying the enrichment in functionally relevant terms (gene ontology, pathways, etc) in chromosomal regions with copy number alterations.

FUNCTIONALITY AND VISUALIZATION

The program

ISACGH (a meta acronym that stands for In Silico Array-CGH) is a web-based integral system that allows studying, within the same context, copy number alterations and gene expression, and provides facilities for the functional profiling of the regions affected. ISACGH can process most of the common gene identifiers and automatically maps them onto chromosome coordinates (human or mouse are available). ISACGH can input gene expression values, genomic hybridization values or both simultaneously. It is not necessarily to use the same platform for chromosomal and expression hybridizations. For example, a case in which a BAC array is used for copy number analysis and a cDNA array is used for gene expression analysis can be analyzed. In principle the number of probes that can be handled depends mainly on the browser used and the memory of the client computer. Current browsers can easily handle high density arrays in the order of 100 000 probes or even more.

Input format

The input format is the one used by GEPAS (19,20) and other similar tools and consists of a tab-delimited text file where the first column correspond to the probe identifiers. The following column(s) correspond to the hybridization intensities (or ratios if two-colour microarrays are used) obtained for each probe in the microarray(s) analyzed. Either genomic hybridizations or mRNA-derived hybridizations are input in the same format. Additionally a file with the chromosomal coordinates of the probes in the chromosomes can be provided. Again, this is a tab-delimited text file with four columns: the fist one contains the probe identifiers, the second one the chromosome in which these are located and the third and fourth ones the chromosome coordinates of the 5′ and 3′ ends of the probes.

Functionality and representation of the results

When genomic hybridization is used, the program predicts the regions with copy number alterations. If only gene expression values are provided, these are mapped onto their chromosomal coordinates. When both, genomic and gene expression values are provided, changes in genomic copy number are predicted and plotted in the same figure together with expression values. Figure 1 shows a combined plot of copy number estimation (blue line) and gene expression (grey bars) in the human chromosome 18. An important aspect is the assessment of the effect of copy number in the global expression of the genes contained in the amplified/lost region. To this end a Student *t*-test has been implemented to assess differential expression between the genes with normal copy number (those in the base line block) and the genes found in regions with copy number alterations. In addition, plots for the direct visualization of the relationship between both expression and copy number can be obtained. Interestingly, if expression values are entered instead of genomic hybridization values, the program can find regions of increased gene expression (RIDGEs) (15).

There are different possibilities for the representation of the results which include several types of multiple-view plots (all the chromosomes of one sample or the same chromosome for multiple samples). In addition, plots of piled samples to detect minimal regions with deletions (or amplifications) in the chromosomes can be obtained. All the results obtained can be visualized in detail in the ISACGH internal viewer but, as an additional and novel

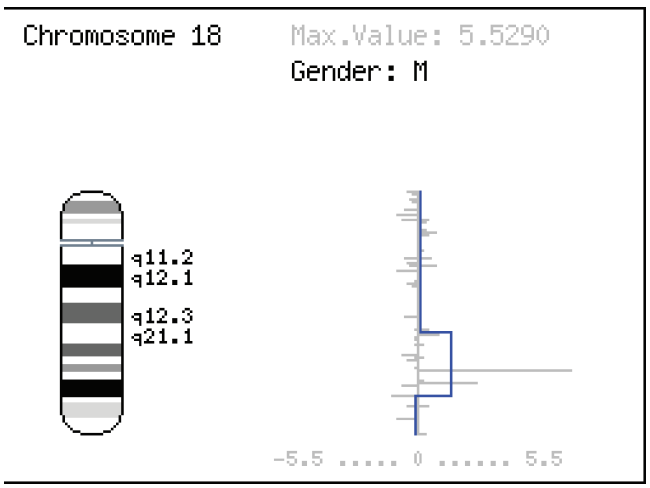


Figure 1. Human chromosome 18. Multiple myeloma (mm) cell line SK-MM-2 (see text) with copy number estimation (blue line) and gene expression values (grey bars). The isowindow segmentation method was used to estimate significant alterations in copy number.

feature, they can also be visualized onto the Ensembl browser.

The distributed annotation system (DAS) is a client-server system in which a single client, in this case the Ensembl (http://www.ensembl.org), integrates information from multiple servers (see http://www.biodas.org). Using the DAS architecture, the Ensembl gathers genome annotation information from multiple distant web sites, collate such information, and display it to the user in its viewer together with the own ensemble data and predictions. Thus, the use of DAS servers for visualization of any genomic feature on the Ensembl viewer offer an excellent environment for the study of the results produced by ISACGH in the genomic context, with the possibility of accessing to any type of available information.

Then, if the Ensembl DAS server option is selected, clicking onto a chromosomal region will produce the creation of a DAS server with information about the probes in the region and the copy number estimation. This information is exported to the Ensembl viewer, which acts as DAS client. Figure 2B shows approximately the same chromosomal region than Figure 2A, but represented in the Ensembl environment. Any genomic feature available in Ensembl in the same chromosomal region can be visualized together with the ISACGH results.

Breakpoint detection

Two methods for breakpoint detection, GLAD (23) and CBS (24), which are among the best performers (11) have been included in the program. We have also developed and included two new methods: a segmentation method (isowindow) and a method based on the slopes of regression in local intervals for copy number change detection. A comparison of the relative performances of the methods implemented was carried out by means of simulated data sets. The new methods proposed here perform at least as well as the GLAD and CBS in terms of tolerance to noise and accuracy in the determination of breakpoints but are more efficient in terms of runtimes (data available in http://bioinfo.cipf.es/downloads/).

Functional profiling of regions with copy number alterations

As previously commented, the ultimate aim of an Array-CGH experiment is to find a molecular explanation for the effects of the detected copy number alterations. The interpretation of genome-scale data is usually performed in two steps: in a first step, genes of interest are selected in this case because they are located in the amplified (or lost) region detected. In a second step, the selected genes of interest are compared to the background (here the rest of genes in the chromosome) in order to find enrichment in any functional category (gene ontology, KEGG pathways, etc.) This comparison to the background is required because otherwise the significance of a proportion (even if high) cannot be determined. Different approaches have been developed to this end (25). Here we will use the FatiGO (22) method, which uses a Fisher's exact test to determine the enrichment in different functional categories. In this case we will analyse the enrichment in GO terms but other functional categories such as KEGG pathways, Interpro functional motifs, Swissprot keywords and some regulatory elements as transcription factor binding sites or other regulatory motifs can also be analyzed with this tool.

A CASE STUDY OF MULTIPLE MYELOMA

To illustrate the concept of functional profiling in the context of array CGH we will use an example of multiple myeloma (MM), an incurable form of haematological neoplasia. The data and the experimental steps followed are described in (26). The aim here was to identify any possible region that contained copy number gains (amplifications), to study the expression of the genes included in that particular region and to understand the possible functional consequences of such alterations.

Data from two-colour hybridizations for both nuclear DNA and transcripts were normalized using the corresponding GEPAS (19,20) module DNMAD and redirected to ISACGH from there. The isowindow method, at medium resolution, was used as the option for the estimation of regions with copy number alterations. The aim was to identify the amplified regions (amplicons) and, to localize and identify the genes that are placed at the amplicon limits. The next step involved the determination of the global expression status of the genes included in these amplicons. And the final aim was to understand the functional consequences associated to the alteration of the expression of such genes.

The analysis was focussed in the chromosome 18, where high level amplification and recurrent gains were found by conventional CGH in cell lines or primary patient samples (27). Within this chromosome, a region with a high level of amplification (amplicon) located at the cytoband 18q21 was detected. MM cell line SK-MM-2 showed a well defined amplicon with an altered gene expression profile (Figure 1). Within the limits of the amplified region several genes display higher expression rates (Figure 1).

Functional profiling of the amplicon revealed a significant enrichment in a number of GO terms in the genes contained in such region. Thus, the GO terms *regulation of cellular process* (GO:0050794) and *regulation of physiological process* (GO:0050791) were significantly over-represented in the amplicon (FDR adjusted p-value = 0.0336). Genes annotated with these terms were: BCL2, MALT1, NEDD4L, MBD2, TNFRSF11A and TCF4. Some of them have annotations at more detailed levels in GO, although the number of genes is too small as to produce statistically significant results. For example BCL2 and MALT1 are annotated as *negative regulation of programmed cell death* (GO:0043069). These results show how the amplification is affecting to a group of functionally related genes and allows conjecturing their global implication in the diseased condition.

DISCUSSION

We present ISACGH, a web-based integrated system that allows simultaneously studying copy number alterations

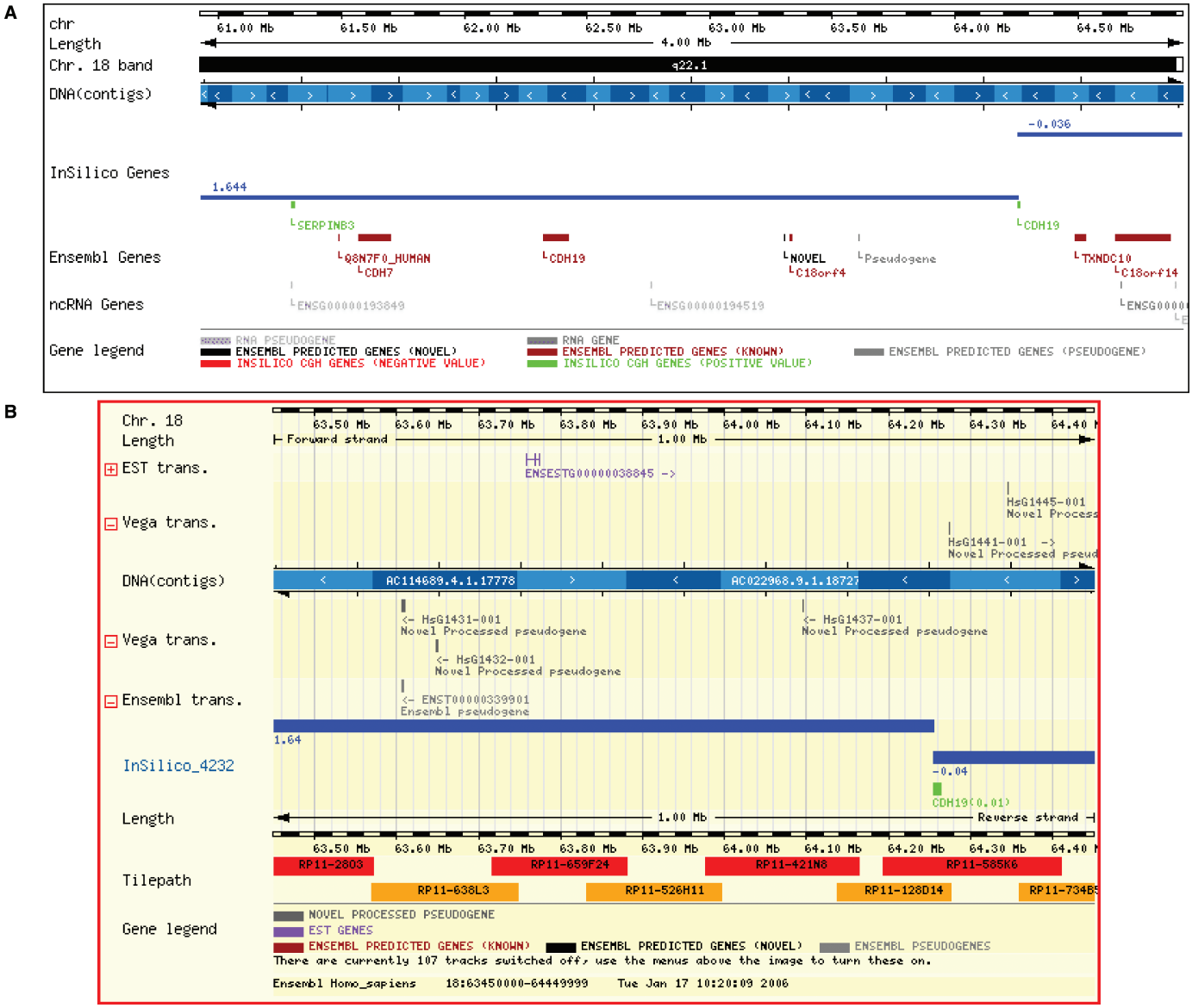


Figure 2. The two zoom options in the breakpoint on the extreme closest to the centromer of the amplicon detected in 18q21.1 in one of the mm cases studied. The two probes form the array shown in the figure (the ones corresponding to *SERPINB3* and *CDH19*) are green because all of them represent amplifications. The blue line represents the copy number estimation. (**A**) ISACGH viewer, (**B**) DAS server.

using array-CGH, their effect on gene expression and the possible functional impact of the chromosomal alteration. In addition, ISACGH is integrated in the GEPAS package, facilitating the normalization, data transformation and other higher-level analysis such as differential gene expression, clustering, etc. This integration may help researchers to overcome the necessity of cumbersome data reformatting operations. Although other two web-based applications for array-CGH data analysis are available [CAPweb (13) and ArrayCyGHt (14)], ISACGH is the only web-based tool offering this combination of analyses to our knowledge.

The results obtained in the case study suggest that the alterations that ultimately lead to MM are not produced by the deregulation of one unique gene, but are rather the combined result of simultaneous deregulations of genes involved in one or more pathways or biological functions. Recent observations on the existence of a non-negligible number of clusters of functionally- related genes suggests that this phenomenon might be more frequent in pathologies characterized by copy number alterations than previously imagined. These findings stress on the importance of the functional profiling for the proper understanding of the functional implications of genomic copy number alterations.

ACKNOWLEDGEMENTS

This work is supported by grants from the Spanish ministry of education and science (BIO 2005-01078) and National Institute of Bioinformatics (www.inab.org) a platform of Genoma España. Funding to pay the

Open Access publication charges for this article was provided by Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Bassett,A.S., Chow,E.W. and Weksberg,R. (2000) Chromosomal abnormalities and schizophrenia. *Am. J. Med. Genet.*, **97**, 45–51.
2. Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.*, **12**, R145–R152.
3. Kallioniemi,A., Kallioniemi,O.P., Sudar,D., Rutovitz,D., Gray,J.W., Waldman,F. and Pinkel,D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.
4. Mantripragada,K.K., Buckley,P.G., de Stahl,T.D. and Dumanski,J.P. (2004) Genomic microarrays in the spotlight. *Trends Genet.*, **20**, 87–94.
5. Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**(Suppl), S11–S17.
6. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, **23**, 41–46.
7. Carvalho,B., Ouwerkerk,E., Meijer,G.A. and Ylstra,B. (2004) High resolution microarray comparative genomic hybridization analysis using spotted oligonucleotides. *J. Clin. Pathol.*, **57**, 644–646.
8. Zhou,X., Mok,S.C., Chen,Z., Li,Y. and Wong,D.T. (2004) Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. *Hum. Genet.*, **115**, 327–330.
9. Hyman,E., Kauraniemi,P., Hautaniemi,S., Wolf,M., Mousses,S., Rozenblum,E., Ringner,M., Sauter,G., Monni,O. *et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.
10. Mahlamaki,E.H., Kauraniemi,P., Monni,O., Wolf,M., Hautaniemi,S. and Kallioniemi,A. (2004) High-resolution genomic and expression profiling reveals 105 putative amplification target genes in pancreatic cancer. *Neoplasia*, **6**, 432–439.
11. Lai,W.R., Johnson,M.D., Kucherlapati,R. and Park,P.J. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.
12. Lockwood,W.W., Chari,R., Chi,B. and Lam,W.L. (2006) Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet.*, **14**, 139–148.
13. Liva,S., Hupe,P., Neuvial,P., Brito,I., Viara,E., La Rosa,P. and Barillot,E. (2006) CAPweb: a bioinformatics CGH array analysis platform. *Nucleic Acids Res.*, **34**, W477–W481.
14. Kim,S.Y., Nam,S.W., Lee,S.H., Park,W.S., Yoo,N.J., Lee,J.Y. and Chung,Y.J. (2005) ArrayCyGHt: a web application for analysis and visualization of array-CGH data. *Bioinformatics*, **21**, 2554–2555.
15. Caron,H., van Schaik,B., van der Mee,M., Baas,F., Riggins,G., van Sluis,P., Hermus,M.C., van Asperen,R., Boon,K. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
16. Hurst,L.D., Pal,C. and Lercher,M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, **5**, 299–310.
17. Stransky,N., Vallot,C., Reyal,F., Bernard-Pierrot,I., de Medina,S.G., Segraves,R., de Rycke,Y., Elvin,P., Cassidy,A. *et al.* (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, **38**, 1386–1396.
18. Zhou,Y., Luoh,S.M., Zhang,Y., Watanabe,C., Wu,T.D., Ostland,M., Wood,W.I. and Zhang,Z. (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.*, **63**, 5781–5784.
19. Herrero,J., Al-Shahrour,F., Diaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
20. Montaner,D., Tarraga,J., Huerta-Cepas,J., Burguet,J., Vaquerizas,J.M., Conde,L., Minguez,P., Vera,J., Mukherjee,S. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **34**, W486–W491.
21. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.*, **33**, W460–W464.
22. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
23. Hupe,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.
24. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
25. Dopazo,J. (2006) Functional interpretation of microarray experiments. *Omics*, **10**, 398–410.
26. Largo,C., Alvarez,S., Saez,B., Blesa,D., Martin-Subero,J.I., Gonzalez-Garcia,I., Brieva,J.A., Dopazo,J., Siebert,R. *et al.* (2006) Identification of overexpressed genes in frequently gained/amplified chromosome regions in multiple myeloma. *Haematologica*, **91**, 184–191.
27. Cigudosa,J.C., Rao,P.H., Calasanz,M.J., Odero,M.D., Michaeli,J., Jhanwar,S.C. and Chaganti,R.S. (1998) Characterization of nonrandom chromosomal gains and losses in multiple myeloma by comparative genomic hybridization. *Blood*, **91**, 3007–3010.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Functional profiling and gene expression analysis of chromosomal copy number alterations

Lucía Conde[1], David Montaner[1, 2], Jordi Burguet-Castell[1], Joaquín Tárraga[1, 2], Fátima Al-Shahrour[1], and Joaquín Dopazo[1, 2*]

[1]Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain;
[2]Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe (CIPF), Valencia, E-46013, Spain;
Joaquín Dopazo* - Email: jdopazo@cipf.es; * Corresponding author

received January 13, 2007; accepted February 11, 2007; published online April 10, 2007

Abstract:

Contrarily to the traditional view in which only one or a few key genes were supposed to be the causative factors of diseases, we discuss the importance of considering groups of functionally related genes in the study of pathologies characterised by chromosomal copy number alterations. Recent observations have reported the existence of regions in higher eukaryotic chromosomes (including humans) containing genes of related function that show a high degree of co-regulation. Copy number alterations will consequently affect to clusters of functionally related genes, which will be the final causative agents of the diseased phenotype, in many cases. Therefore, we propose that the functional profiling of the regions affected by copy number alterations must be an important aspect to take into account in the understanding of this type of pathologies. To illustrate this, we present an integrated study of DNA copy number variations, gene expression along with the functional profiling of chromosomal regions in a case of multiple myeloma.

Keywords: profile; function; gene expression; chromosomal copy number

Background:

Genomic copy number alterations such as gains or losses of chromosomal regions have been shown to be on the basis of many human pathologies. Classical approaches to characterize these genetic aberrations used comparative genomic hybridisation (CGH), in which genomic DNA was hybridised to metaphase chromosomes. **[1]** Recently, the use of different types of microarrays to directly study genomic variations in DNA copy number is becoming more and more popular. Such massive genomic approaches are known as array comparative genomic hybridisation, or Array CGH. **[2]** These new technologies along with the use of expression arrays allow for a highly accurate characterisation of the dependence of gene expression on alterations in genomic copy number. **[3]**

As in many genome-scale methodologies data analysis and, in particular, the biological interpretation of the results constitutes a well-known bottleneck. Specific problems related to the analysis of Array CGH can be circumscribed mainly to two types: appropriate mapping and visualisation of the data onto the chromosomes, and efficient copy number estimation. This last aspect has been the motivation for a number of analytical approaches recently proposed **[4]**, that can be considered the first generation of algorithms for Array CGH analysis. Obviously, copy number variations are expected to have a strong effect on gene expression. **[5, 6]** Nevertheless, the ultimate aim of studies of copy number chromosomal alterations is to understand what is the effect produced in functional terms. In the classical vision one or a few key genes are the causative factors for the this type of pathologies, and the problem consisted in identifying such genes within the region amplified or deleted. The existence of regions in the chromosomes containing coexpressing genes **[7]** which, in addition, are functionally related has recently been reported even in higher eukaryotes. **[8]** Actually, regional arrangements of genes have found to be regulated not only by copy number alterations but also by different mechanisms such as epigenetic modifications. **[9]** This reinforces the functional role of chromosomal regions including groups of functionally related genes and its possible impact on diseases such as cancer. **[10]** These observations give credence to a new vision in which chromosomal alterations can be causing effects not by altering single key genes but by acting on complete molecular sub-systems such as pathways of functionally related genes. Recently, different approaches have focused on the functional aspects of the results of microarray experiments. **[11, 12]** Nevertheless, the possible functional significance at regional level of copy number alterations has been largely ignored. Here we present a combined approach to the study of copy-number alterations, gene expression and functional profiling, exemplified in a case of multiple myeloma. **[13]**

Methodology:

Functional profiling of Array-CGH experiments under this new perspective would require of three steps: 1) detection of regions with copy number variations (the origin of the disease), 2) detection of regional alterations in gene expression (the causes of the disease) and 3) analysis of enrichment in functional terms in the detected regions (the consequences of the alteration or the functional basis of the disease). While copy number alterations can be detected by means of different methods, alterations in the levels of gene expression are not always easy to be detected using the typical methods (t-test or similar) due several factors such as small sample sizes. For this reason here we will only use plots to visualize the effect of one variable (copy number) into the other one (expression level). The third step, the

functional profiling, becomes then the most important aspect of the analysis given that it will provide a functional explanation of the molecular basis of the disease caused by copy number alterations.

Detection of copy number alterations

We have used a segmentation method which is a variant of the circular binary segmentation method **[14]**, for copy number change detection (isowindow).

The isowindow method tries to identify boundaries between regions with a significant change in the values of intensity of hybridisation of the probes by some consecutive steps. Firstly a t-test is used to determine differences between regions around all possible boundary points. Once all the candidate boundaries have been selected (a liberal p-value is used at this stage) there are sorted from small to high minimum p-values. In a second step the boundary candidates in the list with overlapping neighbourhoods are filtered to obtain a refined list of optimal non-overlapping boundary candidates. All the p-values are recalculated for the redefined neighbourhoods and a more stringent threshold is applied here. Finally, regions at both sides of each boundary candidate are again compared with a t-test. If they are not significantly different in their average hybridisation values, then they are merged as a unique region. Otherwise they define two regions with different copy number value. This is a simple and quick procedure that allows for easily changing from fine to coarse resolution by modifying the thresholds for the p-values.

We have compared isowindow to other two methods for breakpoint detection, GLAD **[15]** and circular binary segmentation (CBS) **[14]**, which are among the best performers. **[4]** In the GLAD method a likelihood function with weights determined adaptively is used to solve the copy number estimation problem locally based on data smoothed. Then, the algorithm finds, for each probe, the maximal neighbourhood in which the local constant assumption holds. Each of the constant pieces of the line define a block of probes with similar copy number among them and different copy number from that of the nearby regions. On the other hand, the CBS method selects firstly a segment of the data (a group of probes that are all consecutively arranged in the genome or in a chromosome). The copy number measures of the probes in that segment are compared to those in the reminder dataset using a t-statistic. Hence, the method can distinguish whether the segment chosen has a copy number that is higher or lower than the overall copy number in the data, assumed to be the normal reference. This scheme is iterated exhaustively for all possible segments in the dataset, spotting those that correspond to regions of altered copy number.

An approximation to the relative performances of the methods used was obtained by means of simulated data sets. Such datasets were generated by means of a piecewise constant function plus random alterations normally distributed with mean value and three different levels for the standard deviation (corresponding to noise levels 0.2, 0.5 and 1). A mean value of 0 would correspond to a normal region, without copy number alterations, while mean values lower and higher would correspond to deletions or amplifications at different degrees, respectively. Amplified and deleted regions of different sizes are randomly situated within the simulated normal chromosome and the methods have to locate them at different noise levels. The method proposed here performs at least as well as the GLAD and CBS (Table 1) while being more efficient in terms of runtimes. Isowindow shows a better performance in finding small amplicons.

	Method		
Noise level	GLAD	Isowindow	CBS
0.2	96.9	100.0	90.6
0.5	40.6	62.5	87.5
1.0	9.4	21.9	21.9

Table 1: Percentage of success in finding copy number alterations in the simulation of the four methods for copy number estimation included in ISACGH

Functional profiling of regions with copy number alterations

The final aim of a Array-CGH experiment is to find a molecular explanation for the effects of the detected copy umber alterations. The interpretation of genome-scale data is usually performed in two steps: in a first step genes of interest are selected in this case because they are located in the amplified (or lost) region detected. In a second step, the selected genes of interest are compared to the background (here the rest of genes in the chromosome) in order to find enrichment in any functional category (gene ontology, KEGG pathways, etc.) This comparison to the background is required because otherwise the significance of a proportion (even if high) cannot be determined. Different approaches have been developed to this end. **[11]** Here we will use the FatiGO+ (16) program, which uses a Fisher's exact test to determine the enrichment in different functional categories including gene ontology, KEEG pathways, Interpro functional motifs, Swissprot keywords and some regulatory elements such as transcription factor binding sites or other regulatory motifs. **[17]**

Discussion:

We have implemented all the described functionalities in a program, ISACGH (an acronym for *In Silico* Array CGH), which is used to illustrate the concept of functional profiling of CGH arrays with an example of multiple myeloma (MM), an incurable form of haematological neoplasia.

Nine MM cell lines were obtained from the DSMZ (Deuche Sammlung von Mikroorganismen und Zelkuturen GmbH, Braunschweig, Germany) and were cultured under recommended conditions. DNA and RNA were extracted using supplier's protocols. Microarray assays were performed using the CNIO OncoChip, which contains 7657 different cDNA clones of cancer related genes. **[18]** CGH experiments onto cDNA arrays and hybridisation were performed as described in **[13]** and quantified using the GenePix Pro 5.0 software (Axon Instruments Inc., Union City, CA). Cy3/Cy5 ratio values were normalized using the DNMAD tool from the GEPAS **[19, 20, 21]** and the resulting data were transformed to log2 ratios. Our purpose was to identify any possible region that contained copy number gains (amplifications), to study the expression of the genes included in that particular region and to understand the possible functional consequences of such alterations.

Using the segmentation method as implemented in the ISACGH we could detect a putative amplicon in the chromosome 18 (which remained undetected with both GLAD and CBS, because of the low density of the array, although the effect would have been the same in a high density arrays with a small amplicon) The figure shows the region (left) and the slight, although appreciable, differences in gene expression levels within the amplicon (right).

A unique feature offered by ISACGH is the possibility of obtaining a functional profile of the detected chromosomal regions. When the amplicon is analysed through the FatiGO+ program **[16, 17]** a number of GO terms arise as over-represented in the genes contained in such region. Thus, the GO terms regulation of cellular process (GO:0050794) and regulation of physiological process (GO:0050791) were significantly over-represented in the amplicon (FDR adjusted p-value= 0.0336). Genes annotated with these terms were: BCL2, MALT1, NEDD4L, MBD2, TNFRSF11A and TCF4. Some of them have annotations at more detailed levels in GO, although the number of genes was too small as to produce statistically significant results. For example BCL2 and MALT1 are annotated as negative regulation of programmed cell death (GO:0043069). These observations suggest that some processes altered, that ultimately lead to diseases, are not produced by the deregulation of one unique gene, but are the combined result of simultaneous deregulations of genes involved in a pathway or a particular biological function. In addition, these findings stress the importance of the use of functional profiling methods for the proper understanding and interpretation of the results of the genome-scale experiments. This unique feature included in ISACGH is of extreme importance since growing evidence suggests the existence of clusters of functionally related genes in the chromosomes **[8]** and the possible impact on diseases such as cancer. **[10]**

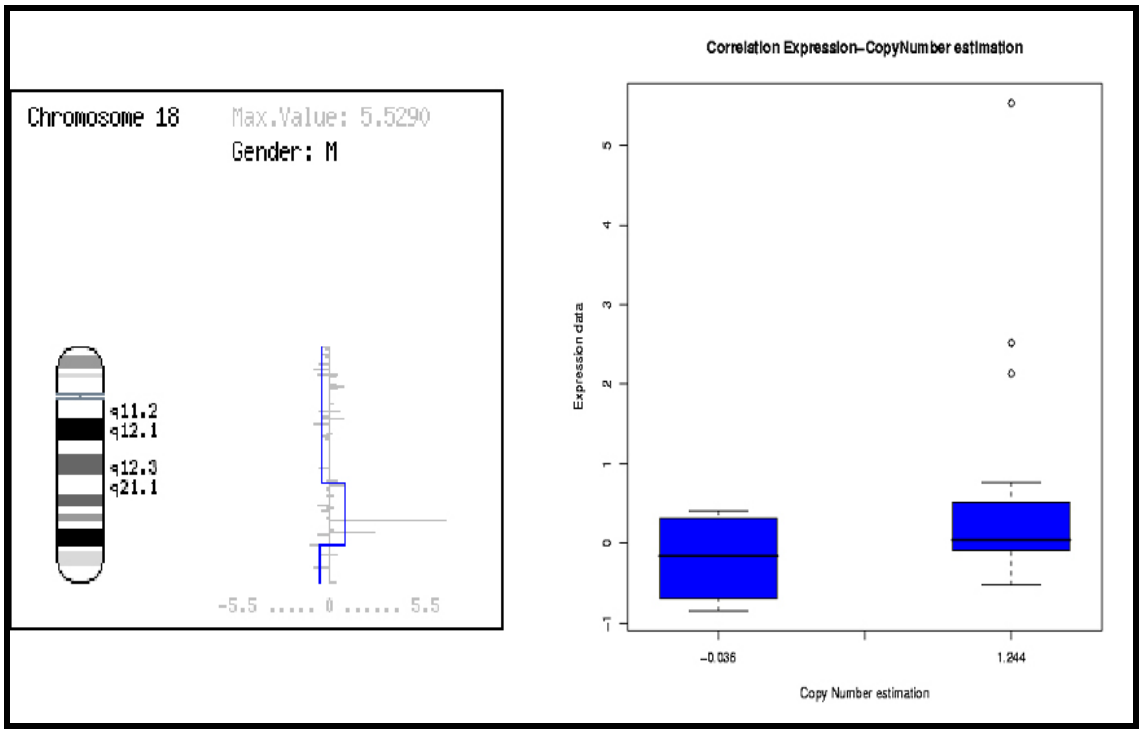


Figure 1: Detection on an amplicon in the chromosome 18 and the relationship between copy number estimation and gene expression. Left: the blue line represents the copy number estimation and the grey bars represent the individual gene expression values represented onto the same chromosomal coordinates. Right: boxplots of gene expression values for the regions with no copy number (a log-ratio of approximately 0) and the amplicon region, which is a duplication (a log-ratio of approximately 1) There is a slight increase in gene expression values in the region of the amplicon

Although ISACGH **[22]** can be used alone, it is tightly integrated in the GEPAS package. **[19, 21, 23]** GEPAS, that stands for Gene Expression Profile Analysis Suite (GEPAS), constitutes one of the most complete resources for microarray data analysis available over the web. GEPAS includes facilities for normalisations, clustering, gene selection, predictors and functional profiling. Thus, different operations (including pre-processing or normalization) can directly be performed within the same environment, without the necessity of any file reformatting step.

Conclusion:

Despite a number of applications dealing with the estimation of genomic copy number have been recently published **[4]**, there are different aspects of the analysis of Array CGH data that have been poorly addressed or even ignored. Recent evidences strongly support the existence of regional arrangements of functionally related genes **[8]**, with obvious consequences for the understanding of diseases characterised by copy number alterations, such as an important number of cancers. **[10]** This fact reduces the validity to the classical vision, in which one or a few key genes would be the causative factors of the disease, and urges to take into consideration the functional dimension in the interpretation of the effects of copy number alterations. In this new scenario, the deregulation of blocks of functionally related genes located in the chromosomal regions with copy number alterations would be behind the disease phenotype.

The methods for functional profiling have proven in many scenarios its usefulness. An obvious challenge is to increase our knowledge in different aspects of function and cooperation between genes in order to be able of applying this methods in a way that allows us to unravel new unknown functional aspects of the biology of the cell and their connections to pathologies.

Acknowledgement:

This work is supported by grants from Fundació La Caixa, NRC Canada-SEPOCT Spain, project BIO 2005-01078 from the MEC and National Institute of Bioinformatics (www.inab.org) a platform of Genoma España.

References:

[01] A. Kallioniemi, *et al.*, *Science*, 258:818 (1992) [PMID:1359641]
[02] D. G. Albertson & D. Pinkel, *Hum Mol Genet.*, 2:R145 (2003) [PMID:12915456]
[03] E. H. Mahlamaki, *et al.*, *Neoplasia*, 6:432 (2004) [PMID:15548351]
[04] W. R. Lai, *et al.*, *Bioinformatics*, 21:3763 (2005) [PMID:16081473]
[05] M. Heidenblad, *et al.*, *Oncogene*, 24:1794 (2005) [PMID:15688027]
[06] D. Pinkel & D. G. Albertson, *Nat Genet.*, 37:S11 (2005) [PMID:15920524]
[07] H. Caron, *et al.*, *Science*, 291:1289 (2001) [PMID:11181992]
[08] L. D. Hurst, *et al.*, *Nat Rev Genet.*, 5:299 (2004) [PMID:15131653]
[09] N. Stransky, *et al.*, *Nat Genet.*, 38:1386 (2006) [PMID:17099711]
[10] Y. Zhou, *et al.*, *Cancer Res.*, 63:5781 (2003) [PMID:14522899]
[11] J. Dopazo, *Omics*, 10:398 (2006) [PMID:17069516]
[12] S. Datta & S. Datta, *BMC Bioinformatics*, 7:397 (2006) [PMID:16945146]
[13] C. Largo, *et al.*, *Haematologica*, 91:184 (2006) [PMID:16461302]
[14] A. B. Olshen, *et al.*, *Biostatistics*, 5:557 (2004) [PMID:15475419]
[15] P. Hupe, *et al.*, *Bioinformatics*, 20:3413 (2004) [PMID:15381628]
[16] F. Al-Shahrour, *et al.*, *Bioinformatics*, 20:578 (2004) [PMID:14990455]
[17] F. Al-Shahrour, *et al.*, *Nucleic Acids Res.*, 33:W460 (2005) [PMID:15980512]
[18] L. Tracey, *et al.*, *Am J Pathol.*, 161:1825 (2002) [PMID:12414529]
[19] http://www.gepas.org
[20] J. Herrero, *et al.*, *Nucleic Acids Res.*, 32:W485 (2004) [PMID:15215434]
[21] D. Montaner, *et al.*, *Nucleic Acids Res.*, 34:W486 (2006) [PMID:16845056]
[22] http://isacgh.bioinfo.cipf.es
[23] J. Herrero, *et al.*, *Nucleic Acids Res.*, 31:3461 (2003) [PMID:12824345]

Edited by Susmita Datta

Citation: Conde *et al.*, Bioinformation 1(10): 432-435 (2007)