Variabilidad transcripcional y enfermedades comunes. Hacia una herramienta diagnóstica

JOSÉ CARLOS RODRÍGUEZ REY

RESUMEN

Las enfermedades comunes, como la diabetes o las enfermedades cardiovasculares, tienen un importante componente genético. El conocimiento de las variantes génicas que las causan es esencial para el diseño de futuras herramientas diagnósticas.

Mientras que las mutaciones localizadas en las regiones codificantes tienen una elevada penetrancia y no pueden explicar la genética de las enfermedades comunes, las variantes en las regiones reguladoras tienen una serie de propiedades que las hacen ideales para hacerlo. En primer lugar, presentan una penetrancia baja. Además, poseen una mayor capacidad de interaccionar con los cambios ambientales.

La existencia de un gran número de polimorfismos en las regiones reguladoras hace necesario implementar estrategias lo más económicas posibles para su identificación y caracterización. La selección de genes candidatos, seguida de análisis bioinformáticos, es una de ellas. Pero el análisis sería incompleto sin la realización de ensayos funcionales cuyos métodos se describen en este capítulo. Los más habituales son el retardo en gel y los ensayos con genes reporteros, una de cuyas principales limitaciones es la imposibilidad de llevarlos a cabo en gran escala. Por esta razón se están diseñando nuevas aproximaciones con este objetivo. La combinación de todos ellos ya ha producido resultados interesantes. Es de esperar que en los próximos años se produzca un desarrollo que permita la aplicación a gran escala y que permita desarrollar algoritmos que sirvan como pruebas para diagnosticar el riesgo de padecer las enfermedades más comunes.

SUMMARY

Common diseases, such as diabetes and cardiovascular disease, have a strong genetic component. A better knowledge of the «pathogenic» variants is needed in order to design future diagnostic tools.

While mutations in coding regions have a high penetrance and therefore cannot explain the genetics of common disease, the variants located in regulatory regions have properties which make then suitable to do it. In the first place, they present low penetrance. Besides, they have the potential to interact with environmental changes.

A large number of polymorphisms in regulatory regions have been described so far. It is desirable the implementation of low-cost methods for their selection and characterization. The selection of candidate genes followed by bioinformatic analysis is one of them. But the functional analysis is still needed. The most common assays, EMSA and those using reporter genes are not high-throughput metods, and therefore new methods are been constantly developed. These methods have already produced some useful results. It is expected that in the next few years new high-throughput methods will be designed and that in the end the results will allow the design of algorithms which can be used as a diagnostic tool for these group of diseases.

INTRODUCCIÓN

Los cambios en la regulación metabólica constituyen la base de una serie de patologías entre las que podríamos destacar la obesidad, la diabetes o las enfermedades cardiovasculares (a partir de ahora las denominaré enfermedades comunes) que tomadas en su conjunto, constituyen la mayor causa de morbilidad y mortalidad en los países de cultura occidental. A pesar de que factores ambientales ligados al estilo de vida, como el tabaquismo, la dieta y la actividad física pueden influir de forma importante sobre su aparición, los estudios familiares indican que alrededor de la mitad del riesgo de padecer estas enfermedades es de origen genético. Si pudiese conocerse cuáles son las variantes génicas que incrementan este riesgo, estaríamos en condiciones de diseñar herramientas diagnósticas que nos ayudarían a la prevención de las mismas.

En un editorial publicado en 1992, Brown y Goldstein sugirieron la necesidad de acabar con lo que ellos denominaron un estado de «*caza de brujas*» genética que se estaba llevando a cabo en el estudio las enfermedades comunes.

Hacían referencia con ello al desorden existente y que consistía en búsquedas a menudo erráticas, en la mayoría de los casos fruto de las preferencias individuales de los investigadores. Como solución propusieron la aplicación de unos postulados similares a los de Koch. Al igual que los microorganismos patógenos, un gen «*culpable*» debería poder ser aislado de individuos afectados (primer postulado). Idealmente, debería poder establecerse una relación de causalidad, reproduciendo la enfermedad en un animal por introducción del gen dañado, gen que posteriormente debería poder recobrarse del animal enfermo (segundo y tercer postulados)(1). Hoy, dieciséis años después, en la base de datos de la Universidad de Cardiff se han anotado ya más de 65.000 mutaciones responsables de enfermedades humanas y, a pesar de su gran importancia social, las enfermedades comunes siguen siendo las grandes desconocidas. A causa de su complejidad las enfermedades comunes no pueden abordarse bajo unos supuestos tan simples. Es difícil hacer estimaciones acerca del número total de genes que intervienen en la aparición de una determinada patología, pero en lo que respecta a las enfermedades cardiovasculares podría haber más de un centenar (2). Y el descubrimiento de las variantes que causan el aumento de susceptibilidad a las mismas sigue constituyendo el gran reto de la era «post-Genoma».

La hipótesis «*Enfermedades comunes-variantes comunes*» propone que las variantes génicas que predisponen a las enfermedades comunes están presentes con una frecuencia elevada en la población general (3). De acuerdo con ella, se podría elaborar un catálogo de polimorfismos comunes y utilizar los estudios de asociación para identificar locus que incrementen la susceptibilidad a padecer una determinada patología. Esta visión sería demasiado optimista en opinión de algunos autores. De acuerdo con ellos, no puede descartarse la existencia de variantes poco frecuentes que determinen la susceptibilidad (4). Aunque la aceptación de uno u otro modelo condiciona en gran medida el diseño de algunos experimentos, hay cosas que son comunes a ambos. Parece claro que las variantes «patológicas» deben identificarse y anotarse. A lo largo de este capítulo se comentará qué tipo de variantes parecen más prometedoras y que métodos son los más idóneos para llevar a cabo su identificación.

ESTRATEGIAS UTILIZADAS PARA LA BÚSQUEDA DE GENES CAUSANTES DE LAS ENFERMEDADES COMUNES

Históricamente, la búsqueda de las variantes genéticas responsables de las enfermedades comunes se ha llevado a cabo de dos formas diferentes. En una de ellas, denominada *estrategia de los genes candidatos*, se seleccionan una se-

rie de genes para su estudio posterior. La búsqueda no es, parafraseando a Goldstein y Brown, una «*caza de brujas*» indiscriminada, sino que parte de un conocimiento biológico previo del papel de los genes y/o del posible mecanismo de los procesos patológicos. Si, por ejemplo, se sabe que el riesgo cardiovascular aumenta en presencia de concentraciones elevadas de triglicéridos y colesterol, o presiones sanguíneas altas, los genes que codifican proteínas que tengan un papel reconocido en estos procesos son candidatos naturales. Hasta ahora, la aplicación de esta estrategia, tanto en su versión de estudios de ligamiento como de asociación, ha sido la que ha producido más y mejores resultados, aunque tiene una limitación que va implícita en su propia naturaleza. Al estar basada en el conocimiento de los procesos, se excluye de entrada a los genes cuya implicación en los mismos es desconocida.

Los problemas debidos al desconocimiento parcial de los procesos patológicos pueden obviarse con el segundo tipo de estrategias: *los barridos genómicos*. Estos consisten en el análisis simultáneo de una serie de marcadores genéticos distribuidos a lo largo del genoma, en grupos de individuos que presentan la patología y su comparación con grupos de control. El barrido genómico permite localizar asociaciones estadísticamente significativas con el fenotipo de estudio en regiones del genoma cuya función ni siquiera es necesario conocer. Como es lógico, la precisión de este tipo de estudios depende del número de marcadores utilizados, y así los estudios iniciales han dado paso a los grandes barridos de la era post-genoma, en la que al conocimiento de la secuencia básica del genoma se une la anotación de millones de polimorfismos de un solo nucleótido (*single-nucleotide polymorphisms o SNPs*), responsables de aproximadamente un 90% de la variabilidad interindividual, y que proporcionan una fuente inagotable de marcadores genéticos (a efectos prácticos se pueden considerar variantes génicas y SNPs como equivalentes). Los experimentos actuales utilizan de forma simultánea cientos de miles de estos últimos colocados en soportes rígidos, los conocidos familiarmente como «*chips*».

Aunque nadie puede negar la utilidad de este tipo de estudios, su relación coste —resultado está siendo cuestionada por los investigadores contrarios a la hipótesis *«Enfermedades comunes— variantes comunes»,* ya que si las enfermedades estuviesen causadas por mutaciones que presentan frecuencias muy bajas, el poder de los análisis estadísticos sería muy pequeño. Además, debido al gran número de variantes estudiadas, se requieren valores de P menores de 5 x 10^{-8} (5). La elección de la hipótesis tiene consecuencias más allá de lo estrictamente académico, puesto que afecta al propio diseño de este tipo de estudios. Esto incluye la elección, tanto del tipo de poblaciones como de la densidad de

los marcadores necesarios para el barrido de una región. Presumiblemente en un futuro, con el abaratamiento progresivo de la obtención de datos de secuencia, este tipo de estudios derive hacia el análisis de la totalidad del genoma lo que, conjugado con un incremento de las capacidades de análisis informático y estadístico, pueda ayudar a superar este debate.

Es importante mencionar que las estrategias de los *genes candidatos* y los *barridos genómicos* son en realidad complementarias. Sobre todo si tenemos en cuenta que los genes localizados en las regiones del genoma que hayan mostrado algún tipo de asociación se convertirán automáticamente en genes candidatos, incluso en ausencia de relaciones funcionales conocidas con el proceso. Además podrían indicar la participación de nuevas rutas o procesos, lo que a su vez puede sugerir el estudio de nuevos genes.

LA IMPORTANCIA DE LAS VARIANTES EN LAS REGIONES REGULADORAS

Se ha comentado en la introducción que la gran mayoría de las variantes génicas causantes de enfermedad caracterizadas hasta ahora son mutaciones que presentan una penetrancia muy elevada (lo que facilita su identificación) y una frecuencia muy baja. Pero además, la mayoría están localizadas en las secuencias del genoma que codifican proteínas. Entre ellas podemos encontrar algunas que contribuyen a la aparición de enfermedades comunes. En el caso particular de las enfermedades cardiovasculares, se pueden destacar las que ocasionan las formas familiares de hipercolesterolemia, causadas por mutaciones del gen del receptor de las lipoproteínas de densidad baja (LDLR) o en el de la apolipoproteína B. Otras formas de enfermedad cardiovascular familiar se han atribuido a mutaciones en el gen de la 5,10-metilen-tetrahidrofolato reductasa (*MTHFR*), de la lamina A/C (*LMNA*), de la «casette» de unión a ATP, subfamilia A gen 1 (*ABCA1*) así como los genes *MEF2A, SCN5A* y *KCNQ1* entre otros (para ver una relación actualizada de los genes implicados el lector puede consultar la referencia (6))

Sin embargo, las formas comunes de enfermedad cardiovascular, y por extensión de todas las enfermedades comunes, no pueden explicarse exclusivamente de acuerdo a una genética mendeliana simple. En primer lugar se trata de procesos en los que interviene un elevado número de genes. Cada uno de estos genes tendría un efecto relativamente pequeño, y combinado bien con otros genes, que podrían actuar como modificadores, bien con determinados factores ambientales, darían lugar a la aparición de los fenotipos que caracterizan a estas patologías.

Pero el análisis más detallado de las enfermedades monogénicas que determinan una patología cardiovascular puede ayudarnos a entender qué tipo de variantes, o mutaciones es necesario buscar. Tomemos de nuevo la HF, una enfermedad monogénica muy estudiada, como ejemplo.

Como ya se ha comentado, la transmisión de la HF es mendeliana y está producida por mutaciones en el gen del receptor de las LDL (LDLR). El número de mutaciones conocido en la actualidad supera el millar. Y la gran mayoría están situadas en la región codificante del gen. Sin embargo, hace unos años se describieron unas familias HF en las que las mutaciones se localizaban en las regiones reguladoras del gen, más concretamente en tres repeticiones presentes en el promotor, que contienen sitios de unión del factor de transcripción Sp1 y de la proteína de unión al elemento de respuesta a esteroles SREBP. Una característica común a estas mutaciones es que la hipercolesterolemia (el rasgo más característico de esta patología) es más suave de lo habitual (7). Incluso, en uno de los casos, descrito por nuestro grupo, ni siquiera todos los portadores del gen presentaban fenotipo HF en ausencia de determinadas condiciones ambientales. Y eso, a pesar de que la mutación producía una reducción de más del 50% en la actividad del promotor (8). Este ejemplo es muy ilustrativo y de él se pueden sacar dos importantes conclusiones. La primera es que los genes que producen fenotipos severos pueden a su vez ser candidatos para explicar las enfermedades que no tienen una transmisión mendeliana. Esto ha sido confirmado recientemente en el caso de la obesidad. Desde hace tiempo se conocen casos de obesidades monogénicas producidas por mutaciones en el gen del receptor de la Melanocortina MC4R. El dato está apoyado por conexiones funcionales puesto que el receptor MCR4 es un regulador de la ingesta, el balance energético y la utilización de la glucosa. Curiosamente, en un estudio muy reciente se ha encontrado una asociación del diámetro de la cintura con marcadores situados en las cercanías del gen que codifica MCR4(9).

La segunda conclusión es que las mutaciones en las regiones reguladoras en general presentan menor penetrancia que las existentes en las regiones codificantes y por tanto podrían ser buenos candidatos para explicar la genética de las enfermedades comunes.

EFECTOS DE LAS MUTACIONES EN LAS REGIONES REGULADORAS

La iniciación de la transcripción es el proceso en el que tiene lugar la regulación genética de forma predominante. La iniciación se produce mediante la colocación de un complejo multi-proteico (complejo de Preiniciación, *PIC*) en

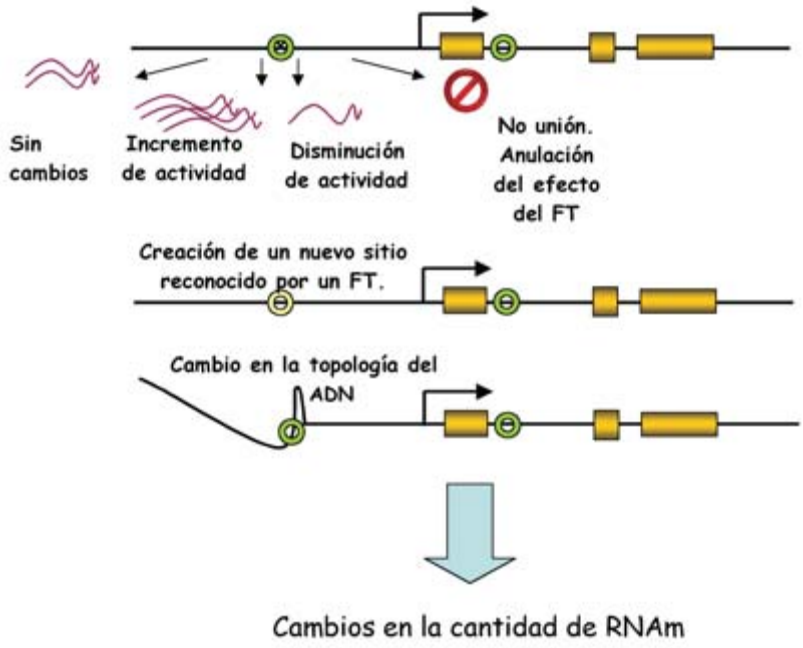

FIGURA 2. *Posible efecto de las variantes en las regiones reguladoras. La existencia de una variante en la región promotora puede dar lugar a incrementos o disminuciones de la actividad del promotor, sea por alteración de la afinidad de los factores de transcripción o por cambios en la topología del ADN. En algunos casos puede crearse un nuevo sitio de reconocimiento de un factor de transcripción, con lo que la nueva variante estará sujeta a un nuevo tipo de regulación.*

un lugar del cromosoma próximo al sitio de iniciación de la transcripción. Para la unión del PIC es necesaria una modificación de las histonas (desacetilación) que permita a las proteínas acceder a la cadena de ácido nucleico. Posteriormente el complejo, bien en pasos, bien previamente preensamblado, se coloca en el promotor para comenzar la síntesis de ARNm (para una discusión de los detalles del proceso el lector puede consultar la referencia (10)). Tanto la accesibilidad de las proteínas al ADN como la estabilidad del complejo formado son los principales factores que determinan la fuerza del promotor y, en definitiva, la cantidad de ARNm sintetizada. En ambos procesos los factores de transcripción desempeñan un papel muy relevante. Llevan a cabo su función uniéndose a sus secuencias de reconocimiento en la región reguladora. Estas secuencias son específicas para cada uno de los factores y los cambios de secuencia pueden dar lugar a cambios en la afinidad del factor o incluso producir la aparición de un nuevo elemento regulador (las posibles consecuencias de los cambios en las regiones reguladoras se resumen en la figura 2). En las secuencias de reco-

nocimiento existe un elevado grado de flexibilidad y muy raramente la secuencia pierde totalmente su capacidad de unir al factor como consecuencia del cambio de un nucleótido. Esta flexibilidad es la que puede servir para explicar la baja penetrancia de las mutaciones y lo que las hace ideales para la explicación de la base genética de las enfermedades comunes.

Además los cambios en las secuencias reguladoras pueden explicar otra característica importante de las enfermedades comunes. En general este tipo de patologías no se hace evidente hasta que concurren una serie de factores ambientales. Por ejemplo, los individuos que van a padecer diabetes tipo II no lo hacen hasta que presentan sobrepeso. En casos como éste es mucho más sencillo explicar la genética de la enfermedad si suponemos un modelo basado en la existencia de una variante en un elemento regulador. Hay que tener en cuenta que no todos los factores de transcripción están activados en la célula en todo momento. De hecho, la síntesis y/o activación de los factores de transcripción es uno de los principales sistemas por los que la célula responde a las señales del ambiente. Por otra parte, algunos factores de transcripción se expresan de forma exclusiva en algunos períodos del desarrollo y otros solamente lo hacen, o son activos, en presencia de determinado estímulo, sea éste químico u hormonal.

A consecuencia de la idoneidad de las regiones reguladoras para explicar la genética de las enfermedades comunes cada vez más grupos dirigen su atención al estudio de estas variantes reguladoras (o SNPs reguladores, o rSNPs). Hasta el momento no existen muchos estudios sistemáticos sobre el efecto funcional de los SNPs, pero los que han sido publicados, sugieren que una parte importante de los SNPs localizados en los promotores producen alteraciones en la expresión del gen, es decir, son verdaderos rSNPs. En uno de estos estudios, llevado a cabo con el objeto de conocer la frecuencia de los rSNPs, se analizaron las 500 bp inmediatamente previas al inicio de la transcripción de 170 genes no relacionados, escogidos de entre aquellos cuyos promotores estaban anotados en la Eukaryotic Promoter Database. Se encontraron mutaciones (SNPs) en aproximadamente un 35% de los genes (hay que considerar que solamente se analizaron las 500 primeras pares de bases). De éstas aproximadamente una tercera parte producían cambios en la transcripción de los genes (11). En otro estudio posterior, realizado por el mismo grupo, se analizaron los promotores de 31 genes que se expresan de forma específica en el cerebro. En ellos se encontraron 60 polimorfismos. Los análisis posteriores indicaron que un 22% de los mismos producían cambios reconocibles en la actividad del promotor (12).

Los dos artículos mencionados demuestran que existe una gran fuente de variabilidad funcional en los elementos reguladores del genoma, lo que resulta-

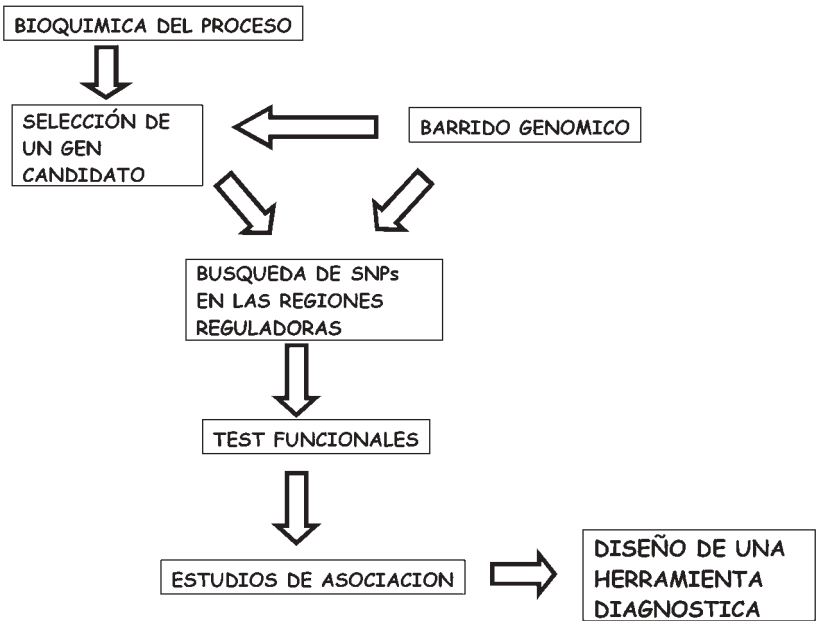

FIGURA 1. *Diagrama de flujo para el diseño de herramientas diagnósticas. La selección de los genes candidatos se llevará a cabo bien por conocimiento previo de los mecanismos que producen la patología, bien por haber sido identificados en estudios genómicos previos. Los SNPs existentes en la regiones reguladoras se seleccionarán a su vez mediante la combinación de estudios bioinformáticos y test funcionales. Los SNPs de interés se utilizarán en estudios de asociación con la enfermedad. Aquellos que hayan resultado positivos servirán a posteriori para el diseño de herramientas diagnósticas.*

ría en una gran plasticidad en cuanto a la diversidad de la expresión genética. Puede discutirse sin embargo que, al tratarse de estudios basados en las técnicas de retardo en gel y expresión de genes reporteros (ver más adelante), en ellos no se analiza que los polimorfismos tengan ningún tipo de asociación con fenotipos concretos y por lo tanto su importancia real sería discutible.

Este aspecto ha sido analizado en un artículo reciente. En él se describe un estudio llevado a cabo sobre los polimorfismos en los promotores de varios genes de gran importancia para el ciclo celular; los genes *CDKN2A, CDKN2B, CDKN1A* y *CDKN1B* que codifican inhibidores de quinasas dependientes de ciclinas. Como era de esperar, todos los promotores estudiados presentan varios polimorfismos y, como también era esperable, varios de estos polimorfismos producían diferencias en la capacidad de unión de factores de transcripción. A continuación se estudió la asociación de estos polimorfismos en un grupo de pacientes de leucemia linfoblástica aguda. La mayoría de los polimorfismos no

mostraron asociaciones estadísticamente significativas con la enfermedad y cuando lo hacían, éstas eran mucho más fuertes en presencia de otro de los polimorfismos en el mismo promotor(13). El hecho de que sean los haplotipos, y no los polimorfismos individuales, los que muestren asociación con una patología es un resultado de importantes consecuencias, ya que indica que el efecto de los polimorfismos individuales es muy pequeño, y solamente en combinación con otros muestran una asociación lo suficientemente fuerte con el fenotipo. Ello previene su detección individual en estudios de asociación convencionales, y desde luego en los barridos genómicos (que como hemos mencionado tienen mucho mayores limitaciones estadísticas). Y sugiere que tal vez la estrategia adecuada debería ser la contraria a la utilizada hasta ahora, y sea necesario anotar las variantes funcionales y, una vez seleccionadas, llevar a cabo los estudios de asociación. O una mixta en que ambos estudios se lleven a cabo simultáneamente (figura 1). En cualquier caso los estudios funcionales parecen ser muy relevantes y a la descripción de los mismos está dedicada la segunda parte de este capítulo.

ESTRATEGIAS BIOINFORMÁTICAS PARA LA SELECCIÓN DE SNPS EN REGIONES REGULADORAS

En la actualidad no existen métodos de gran rendimiento para el análisis funcional de los polimorfismos. Los análisis normalmente se centran en un grupo de genes cuidadosamente seleccionados, en general por poseer una relación funcional: por ejemplo que codifiquen proteínas que intervengan en una misma ruta metabólica, que estén inducidos por la misma sustancia, etc

Tras la selección del grupo de genes candidatos es necesario llevar a cabo la identificación de los rSNPs presentes en las regiones reguladoras. *A priori* es muy difícil saber cuales son las regiones reguladoras del genoma. La cromatina está organizada en dominios independientes separados por elementos aislantes (*insulators*) que se corresponden con dominios funcionales, pero éstos son de gran tamaño y poseen una topología que hace posible que secuencias aparentemente muy distantes puedan estar próximas en el espacio (14), por lo que no puede descartarse la participación de secuencias alejadas en la secuencia en la regulación de un gen. A pesar de ello la mayoría de los estudios sugieren que existe una mayor concentración de elementos reguladores en las secuencias cercanas al promotor por lo que, salvo en el caso de grupos de genes que comparten regulación, el análisis de las 3 kilobases anteriores al sitio de iniciación de la transcripción se considera suficiente.

En los bancos de datos públicos se han anotado ya del orden de 10 millones de SNPs, con lo que la mayoría de los rSNPs más frecuentes deberían estar representados. Sin embargo, ya se ha comentado la posibilidad de que algunas variantes poco frecuentes estén relacionadas con las enfermedades comunes(4). Además, podrían existir variantes «locales», por lo que la obtención de nuevos datos de secuencia en nuestra población de estudio estaría más que justificado.

Los rSNPs se localizan en sitios de unión de factores de transcripción y el cambio de nucleótido que da origen al rSNP debe resultar en un cambio en la capacidad de unir el factor. Como durante el proceso de selección se carece de información funcional, es necesario recurrir a aproximaciones bioinformáticas que puedan indicar con un cierto margen de confianza que a) la secuencia sea capaz de unir factores de transcripción y b) que el cambio de la secuencia produzca un cambio en la fuerza de la unión. Hasta hace relativamente pocos años, los algoritmos diseñados para la identificación de sitios de unión de factores de transcripción funcionaban muy bien para los organismos sencillos como levaduras, pero originaban una gran cantidad de falsos positivos en mamíferos. Afortunadamente en los últimos años han aparecido nuevos algoritmos muchos más fiables(15) Incluso, gracias a la disponibilidad de los datos de secuencia de varios genomas de mamíferos, algunos grupos han diseñado herramientas informáticas que permiten no solo identificar los sitios de unión de factores de transcripción sino la predicción del efecto sobre la transcripción de un gen de la presencia de un SNP en las secuencias reguladoras. Este sistema se ha utilizado con relativo éxito en la determinación de elementos polimórficos en los promotores de genes que codifican proteínas involucradas en procesos de defensa contra la oxidación y de enzimas detoxificantes(16)

Para incrementar la probabilidad de identificar sitios de unión reales se utiliza lo que se ha dado en llamar la «*huella dactilar filogenética*». El método asume que, ya que las secuencias funcionales tienen un mayor grado de conservación inter-especie, aquellas secuencias de la región promotora que poseen un grado de conservación elevado tienen una mayor probabilidad de contener elementos reguladores. Así, basta una simple comparación de las regiones 5' de un gen entre varias especies de mamíferos para poder llevar a cabo esta selección. Aunque este método funciona muy bien, y generalmente se acepta como un sistema de identificación de elementos reguladores conservados, (el resultado de uno de éstos análisis se muestra en la figura 3) no puede olvidarse que algunas veces las secuencias de los primates no presentan una homología muy elevada con las secuencias de los roedores (17). Además, existen secuencias ex-

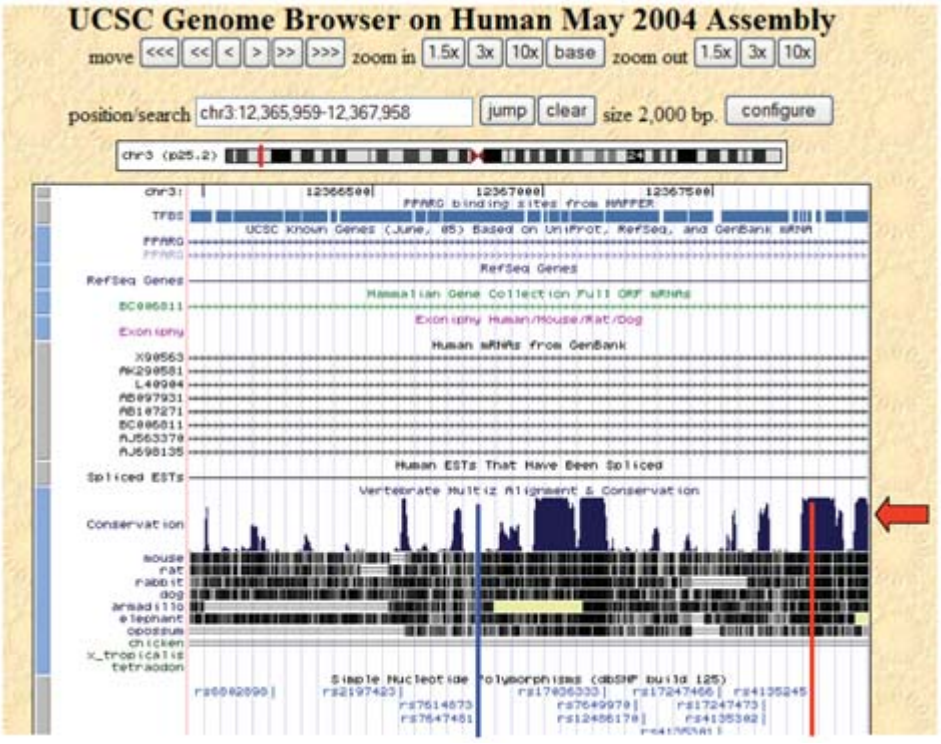

Figura 3. *Pantalla típica de uno de los buscadores utilizados para la selección bioinformática de SNPs en regiones reguladoras. En este caso se trata del promotor del gen que codifica el factor de transcripción PPAR Á. La localización de los SNPs se indica por una pequeña barra vertical al lado del nombre de los mismos (rs…). Los grandes bloques de color azul (flecha roja) señalan las zonas de conservación de secuencia entre las diferentes especies. Como ejemplo se muestran la posición de un SNP en una región no conservada (línea azul vertical gruesa) o conservada (línea roja vertical gruesa). Este último SNP (rs413545) sería un buen candidato para posteriores estudios funcionales.*

clusivas del genoma humano (*Human Accelerated Regions* o HARs) alguna de las cuales puede actuar como un elemento regulador (Pampín y Rodríguez-Rey, resultados sin publicar). Por lo tanto, los métodos filogenéticos deben ser utilizados con una cierta precaución y con el mayor conocimiento posible de las rutas en los que los genes se hallan implicados. Esto puede constituir un problema en el caso de los genes candidatos obtenidos por barridos genómicos, ya que en muchos casos la función de estos genes es desconocida. Esta dificultad podría solventarse en un futuro no muy lejano con el desarrollo de los paneles de ARN de interferencia (18)

MÉTODOS PARA EL ANÁLISIS FUNCIONAL DE SNPS EN LAS REGIONES REGULADORAS. MÉTODOS «IN VITRO»

Para confirmar la existencia de posibles rSNPs, el cambio que producen debe confirmarse mediante un análisis funcional. Como se ha comentado, la gran mayoría de los rSNPs producen su efecto mediante cambios en las secuencias de unión de factores de transcripción (para una discusión de cómo un SNP puede también afectar a la transcripción cambiando la topología del ADN ver (19)), lo que a su vez se traducirá en cambios en la transcripción de los genes en cuyas secuencias reguladoras están situados. Por eso, uno de los métodos más importantes para confirmarlo es el estudio de unión a proteínas por retardo en gel, más conocido por sus siglas inglesas EMSA (Electrophoretic Mobility Shift Assay). El método se basa en que la unión de una proteína (un factor de transcripción en este caso) al ADN (su sitio de reconocimiento) da lugar a una nueva especie molecular de mayor tamaño y con una movilidad electroforética reducida. En un gel podrá verse como una banda que se retrasa con respecto a la banda correspondiente al ADN libre.

Históricamente, el EMSA ha sido el método más utilizado para la identificación de secuencias de unión de factores de transcripción. Se trata de un método relativamente sencillo, que utiliza un oligonucleótido de doble cadena de una longitud de 20 a 30 nucleótidos que contiene la secuencia en la que el posible rSNP ha sido localizado. El fragmento marcado, mediante métodos radiactivos o fluorescentes, se mezcla con un extracto nuclear de células en las que el gen se exprese, lo que da lugar a la formación del complejo ADN— proteínas. La mezcla se somete a una electroforesis en geles de poliacrilamida, en donde la disociación del complejo no se produce por efecto de la menor fuerza iónica del tampón, así como por el efecto «jaula» producido por la propia matriz del gel. El complejo tiene una menor movilidad que el ADN libre, lo que permite su identificación. Si la presencia de un rSNP provoca un cambio de afinidad se observará un cambio en la intensidad o en la localización de la banda. La cuantificación de los parámetros de unión se pueden determinar mediante competición con el oligonucleótido sin marcar (figura 4a).

Si conocemos qué factor de transcripción se une a la secuencia, se puede añadir un anticuerpo específico a la reacción. La unión del anticuerpo al FT presente en el complejo daría lugar a un complejo de mayor tamaño, que sufriría un retraso mayor (super-retardo, figura 4b).

El EMSA es un método sencillo y muy sensible y con él se han determinado la mayoría de los sitios de unión de factores de transcripción conocidos,

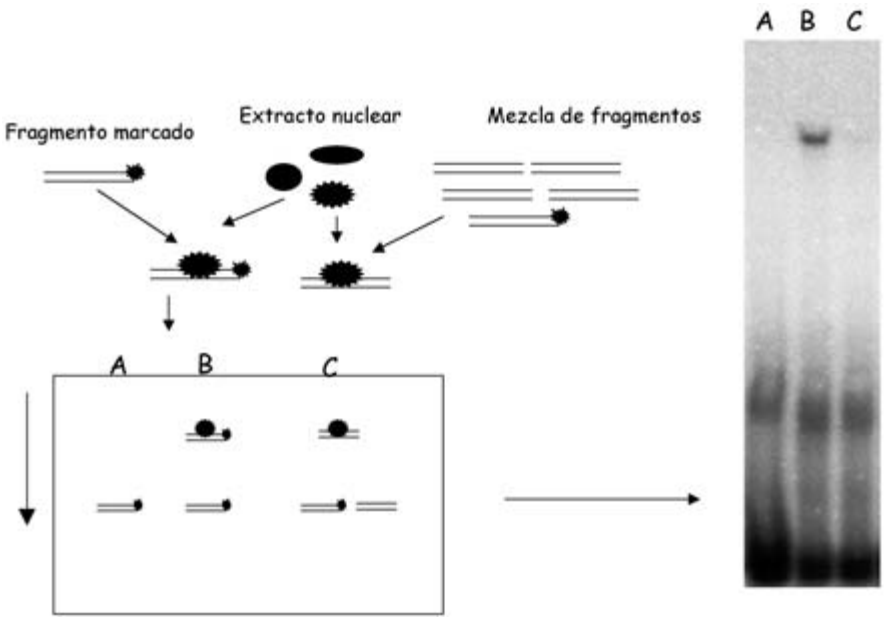

Figura 4a. *Esquema de un experimento de retardo en gel (EMSA). Un fragmento de ADN de doble cadena marcado radiactivamente (calle A) se incuba con un extracto de proteínas. El complejo resultante aparece como una banda nueva en la calle B. la comprobación de la especificidad se lleva a cabo compitiendo con un exceso del mismo oligonucleótido sin marcar, que desplaza del complejo al oligonucleótido marcado. Como consecuencia la banda retardada desaparece del gel (calle C).*

pero no está exento de inconvenientes. El primero es que no puede hacerse con un gran número de muestras de forma simultánea. Además, en ocasiones no puede considerarse cuantitativo, debido a problemas como el ruido de fondo o la degradación de las bandas. Cuando las diferencias son muy pequeñas sería deseable contar con métodos con mayor sensibilidad y adaptables. El FRET (Förster Resonance Energy Transfer) parece ser uno de ellos. Se basa en que cuando dos fluoróforos están cercanos (a menos de 100 Å de distancia) se produce un cambio de fluorescencia cuantificable. EL FRET ya se ha utilizado para la cuantificación de uniones ADN-proteína y se espera una mayor utilización en los años venideros (20). También se están aplicando técnicas basadas en ELISA, en el que los oligonucleótidos inmovilizados se incuban con proteína recombinante o extractos nucleares y se revelan con un anticuerpo específico para la proteína de interés. Para un mayor rendimiento, utilizando la incorporación de oligos inmovilizados en una plataforma, el método se podría adaptar para ser capaz de evaluar cientos de sitios de unión en un solo experimento (21).

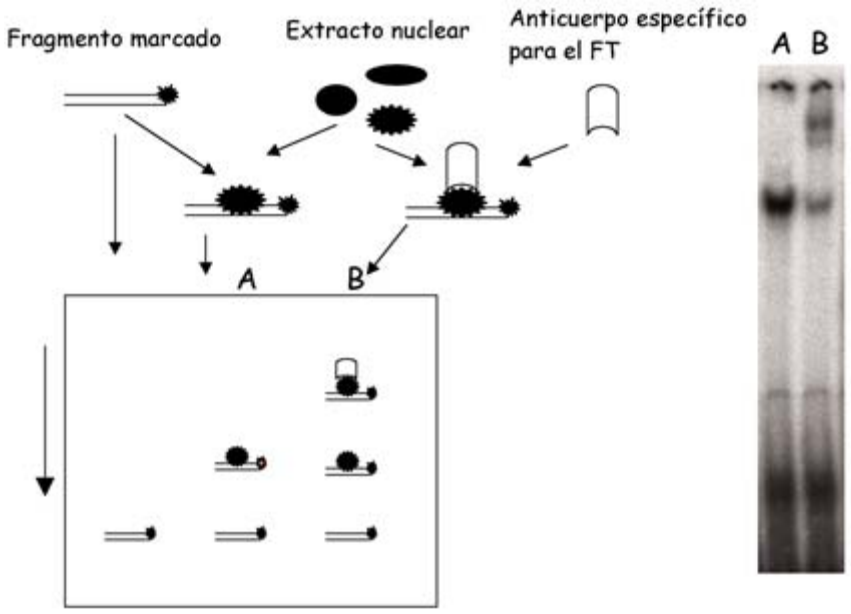

FIGURA 4B. *Identificación de la proteína en un complejo mediante la utilización de anticuerpos específicos. Las calles A y B son equivalentes a las de la figura a. La adición de un anticuerpo específico a la reacción produce un complejo de mayor tamaño que da lugar a una banda con respecto al complejo inicial. Las figuras de los geles son de (33).*

INMUNOPRECIPITACIÓN DE LA CROMATINA (CHIP)

Aunque los métodos de ensayos de unión *in vitro* han demostrado ser extraordinariamente útiles, está claro que sus resultados no dejan de ser algo artificiales. Por una parte, la estructura de un oligonucleótido tiene poco en común con la verdadera estructura del ADN existente en la cromatina. Además, para una unión óptima podría ser necesaria la presencia de componentes celulares ausentes de los extractos. Por último, el hecho de que un extracto contenga una proteína capaz de unirse al ADN no quiere decir que esté efectivamente unida en el momento en el que el proceso se está estudiando. La inmunoprecipitación de la cromatina (ChIP) es un ensayo que permite una aproximación al estado endógeno de la célula, algo así como una instantánea del proceso de unión.

En esta técnica las uniones proteína-ADN se fijan utilizando formaldehido. A continuación el ADN se rompe por sonicación y los fragmentos que están unidos al factor de transcripción se precipitan con un anticuerpo específico contra el mismo. Para identificar la secuencia a la que el factor de transcripción se en-

cuentra unida se han utilizado diferentes métodos que van desde métodos de hibridación clásicos como *Southern blot* o incluso el clonaje de los fragmentos seguido de la determinación de la secuencia, pero en la actualidad la PCR es el método de elección. El método además puede escalarse para el estudio de interacciones a lo largo de todo el genoma, utilizándose en este caso la hibridación a un chip que contenga miles de muestras diferentes (ChIP-on-chip (22)).

MÉTODOS PARA EL ESTUDIO DE LA INFLUENCIA DE LOS SNPS EN LA EXPRESIÓN DE LOS GENES

Los experimentos de unión de las proteínas al ADN dan mucha información acerca de cómo la presencia de un SNP en la región reguladora altera las propiedades de unión de la secuencia. Sin embargo, no puede olvidarse que el efecto final de un cambio de las propiedades una unión de un factor de transcripción es un cambio en los niveles de expresión de un gen. Por ello, para una confirmación del efecto de un SNP hay que recurrir a los ensayos de expresión.

El ensayo con genes reporteros ha sido, y sigue siendo, el más utilizado para el estudio de la fuerza de los promotores. El promotor objeto de estudio es clonado directamente delante de un gen que codifique una proteína fácilmente cuantificable (gen reportero) en un vector que carezca de otro promotor para la expresión de ese gen. El plásmido se introduce por transfección a células en cultivo en donde expresará la proteína codificada por el gen reportero. La cuantificación de la actividad de la proteína (o su cantidad) se hace antes de las 72 horas posteriores a la transfección y sus valores permiten estimar de forma bastante exacta la actividad del promotor o de la región reguladora. La Cloranfenicol Acetil Transferasa (CAT), un enzima responsable de la resistencia al Cloranfenicol, fue el gen reportero por excelencia hasta hace algunos años(23), pero ha sido desplazado por el gen de la luciferasa de *Photinus pyralis*, cuya actividad puede cuantificarse con un luminómetro y se mantiene lineal en un rango de cinco a seis órdenes de magnitud (24).

La eficiencia de la transfección es el factor limitante del proceso y el método utilizado depende del tipo celular. La formación de precipitados de complejos de ADN con fosfato cálcico (25) ha dado paso a técnicas mucho más eficientes. Para células que crecen en suspensión, funcionan muy bien los métodos basados en la formación de poros en la membrana, cuya formación se induce por la aplicación de un choque eléctrico(26). Para las células que crecen en monocapas, existen en el mercado numerosos lípidos catiónicos que forman comple-

jos con el ADN y que, al fusionarse con la membrana, proveen al ADN de una vía de entrada al interior de la célula(27). Sea cual sea el método utilizado, la eficiencia de la transfección debe ser corregida en cada experimento. Normalmente, esto se consigue mediante la co-transfección de otro plásmido en el que un diferente gen reportero se clona delante de un promotor de fuerza conocida (el promotor temprano-intermedio del citomegalovirus y el promotor temprano del virus SV40 son los más utilizados). En principio, cualquier gen reportero puede ser válido, siempre que sea diferente al que porta el plásmido problema, pero el gen luciferasa de *Renilla reniformis* es el más utilizado en la actualidad. Como esta luciferasa utiliza un sustrato diferente de la de *Photinus*, su cuantificación puede llevarse a cabo en el mismo ensayo sin necesidad de dividir el extracto.

MÉTODOS *IN VIVO* PARA LA CUANTIFICACIÓN DE LA EXPRESIÓN ESPECÍFICA DE ALELO

Los genes reporteros permiten determinar el efecto de un SNP sobre la transcripción. Pero, al igual que ocurre con el EMSA, las regiones reguladoras se encuentran en un ambiente no natural. Por ello es deseable poder analizar el efecto de los polimorfismos en la propia célula. Para lograrlo se han diseñado diferentes métodos.

El primero de los métodos consiste en la extensión de un cebador con un único nucleótido (single-nucleotide primer extension, SNuPE). Diseñado originalmente para la detección de alelos mutantes, su utilidad para medir la expresión específica de alelos se comprobó con el caso más extremo de ésta: la impronta génica. En el SNuPE, los transcritos de ambos alelos se amplifican con cebadores que flanquean el SNP. Para determinar la cantidad relativa de cada uno de los transcritos, un oligonucleótido cuyo extremo 3' está localizado justo delante del SNP se extiende en dos reacciones independientes, utilizando en cada una de ellas el nucleótido marcado correspondiente a cada uno de los alelos. La relación de la radiactividad incorporada en cada una de las reacciones refleja la abundancia del ARNm de cada uno de ellos(28). Una de las grandes ventajas del método es la posibilidad de llevar a cabo experimentos a gran escala mediante el acoplamiento del mismo con chips de DNA(29). Se ha empleado con éxito en un estudio llevado a cabo en los linfoblastos de miembros de14 familias en los que se determinó la expresión de 3554 genes. El estudio reveló que los patrones de expresión génica son heredables, proporcionando así una importante confirmación de la importancia de las pequeñas diferencias de expresión en la determinación de los fenotipos complejos(30).

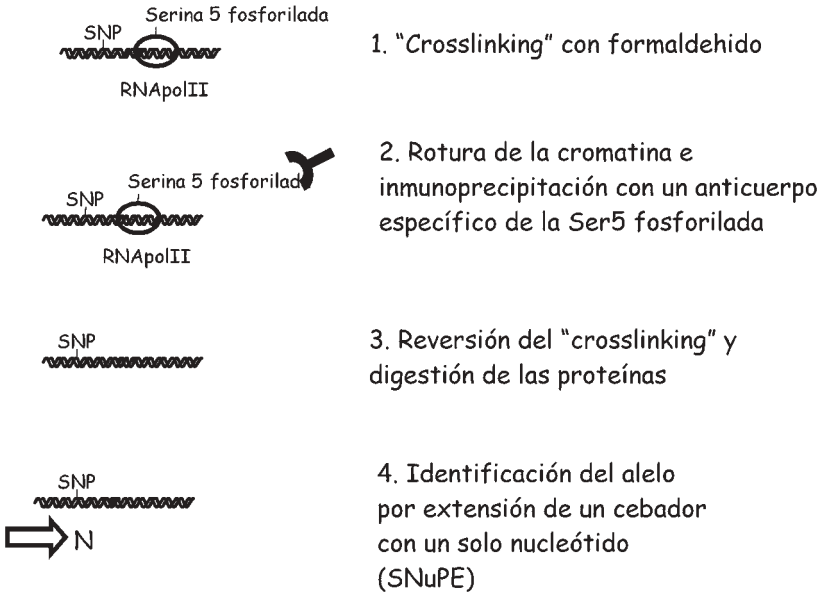

FIGURA 5. *Utilización de la inmunoprecipitación de la cromatina para la determinación de la expresión específica de alelo (31). En los genes que se están transcribiendo de forma activa, la RNA polimerasa se encuentra fosforilada en el residuo de Serina 5 de su dominio carboxi-terminal. Tras fijar las proteínas al ADN mediante formaldehido, el ADN se fragmentará por sonicación. Los fragmentos unidos a la polimerasa activa se inmunoprecipitarán con un anticuerpo específico para la fosfoserina 5 por lo que el precipitado contendrá todos aquellos fragmentos que se están transcribiendo. Tras liberar el ADN, se procederá a la identificación del alelo que se expresa mediante un método de extensión con nucleótidos marcados (SNuPE, ver texto).*

Puesto que están basados en la amplificación de los transcritos, los métodos basados en SNuPe no pueden discriminar en el caso de que los SNPs estén fuera de ellos. Esta es una limitación muy importante si consideramos que excluye todas las secuencias reguladoras y, por tanto, a todos los rSNPs. Se han arbitrado algunas soluciones al problema. Un elegante método diseñado por Knight y cols.(31) permite inmuno-precipitar todos los genes cuya expresión se está llevando a cabo. Para ello se utiliza un anticuerpo contra la fosfoserina 5 del dominio carboxi-terminal de la RNA polimerasa II. La fosforilación de esta Serina tiene lugar durante el proceso de elongación del RNA y por tanto solamente está presente en genes cuya transcripción es activa (figura 5).

Otra limitación importante de los métodos de análisis de expresión específica de alelo tiene que ver con la existencia de más de un SNP en un promotor. Así, en el caso de que pueda demostrarse la existencia de diferencias de expresión entre las dos copias del gen no puede saberse cuál de los SNPs es el responsable de

los cambios. En estos casos haría falta la confirmación de los efectos con métodos en los que los SNPs puedan estudiarse de forma independiente, como es el caso de los estudios *in vitro* o con genes reporteros. De ahí que una combinación de los métodos descritos con anterioridad parece ser la solución más razonable.

CONCLUSIONES

El mapeo de variantes génicas que contribuyen a las enfermedades complejas constituye uno de los mayores retos de la era post-genoma. Entre estas variantes existentes, las situadas en las regiones reguladoras están recibiendo mucha atención en la actualidad puesto que su baja penetrancia podría explicar mejor la genética de las enfermedades complejas. La identificación de los rSNPs que producen cambios funcionales en la expresión del gen es, a día de hoy, un proceso largo y tedioso, dado el gran número de variantes existentes. Como consecuencia, la mayoría de los estudios llevados a cabo hasta ahora se han centrado en el estudio de un número pequeño de genes. La selección de los genes basada en los barridos genómicos o en un buen conocimiento de la Bioquímica, unida a una buena selección bioinformática y a la utilización de sencillos métodos *in vitro* está permitiendo la caracterización de un número ya importante de rSNPs que serán de utilidad en posteriores estudios. Sirva como ejemplo el estudio reciente de los promotores de 176 genes que codifican receptores acoplados a proteínas G en el que la combinación de estos métodos ha permitido determinar la existencia de 20 rSNPs con importantes efectos sobre la transcripción de estos genes (32)o el llevado cabo por nuestro grupo para caracterizar más de 50 rSNPs en promotores de genes importantes del metabolismo lipídico (Pampín y cols, en preparación). Es más que necesario desarrollar técnicas que permitan escalar este tipo de análisis y adaptarlas a la cantidad de datos que pueden obtenerse ya con la obtención de secuencias a gran escala. Es probable que en unos años estemos en situación de hacerlo y con ello de obtener un perfil de rSNPs, que constituirá una poderosa herramienta diagnóstica que permita predecir el riesgo de padecer las enfermedades que constituyen los principales problemas sanitarios en la actualidad.

AGRADECIMIENTOS

La investigación propia mencionada ha sido realizada con varias ayudas del Fondo de Investigaciones Sanitarias del Instituto de Salud Carlos III. El autor quiere agradecer a Sandra Pampín su ayuda en la realización de la figura 2.

ABREVIATURAS

ChIP, Chromatin Immunoprecpitation; EMSA, Electrophoretic Mobility Shift Assay; FRET, Förster resonance energy transfer; FT, Factor de transcripción; HF, Hipercolesterolemia familiar; LDL, Lipoproteínas de baja densidad; PCR, Polymerase - Chain Reaction; rSNP, Regulatory Single-Nucleotide Polymorphism; SNP, Single-Nucleotide Polymorphism; SNuPE, Single - Nucleotide Primer Extension.

BIBLIOGRAFÍA

1. Brown, M. S., y Goldstein, J. L. (1992) Koch's postulates for cholesterol. *Cell* **71**, 187-188.
2. Lusis, A. J. (2003) Genetic factors in cardiovascular disease. 10 questions. *Trends in cardiovascular medicine* **13**, 309-316.
3. Lander, E. S. (1996) The new genomics: global views of biology. *Science (New York, N.Y* **274**, 536-539.
4. Pritchard, J. K. (2001) Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics* **69**, 124-137.
5. Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., y Abecasis, G. R. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature genetics* **40**, 161-169.
6. Arnett, D. K. (2007) Summary of the American Heart Association's scientific statement on the relevance of genetics and genomics for prevention and treatment of cardiovascular disease. *Arteriosclerosis, thrombosis, and vascular biology* **27**, 1682-1686.
7. Koivisto, U. M., Palvimo, J. J., Janne, O. A., y Kontula, K. (1994) A single-base substitution in the proximal Sp1 site of the human low density lipoprotein receptor promoter as a cause of heterozygous familial hypercholesterolemia. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 10526-10530.

8. Mozas, P., Galetto, R., Albajar, M., Ros, E., Pocovi, M., y Rodriguez-Rey, J. C. (2002) A mutation (-49C>T) in the promoter of the low density lipoprotein receptor gene associated with familial hypercholesterolemia. *Journal of lipid research* **43**, 13-18.

9. Chambers, J. C., Elliott, P., Zabaneh, D., Zhang, W., Li, Y., Froguel, P., Balding, D., Scott, J., y Kooner, J. S. (2008) Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nature genetics* **40**, 716-718.

10. Lemon, B., y Tjian, R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes & development* **14**, 2551-2569.

11. Hoogendoorn, B., Coleman, S. L., Guy, C. A., Smith, K., Bowen, T., Buckland, P. R., y O'Donovan, M. C. (2003) Functional analysis of human promoter polymorphisms. *Human molecular genetics* **12**, 2249-2254.

12. Buckland, P. R., Hoogendoorn, B., Guy, C. A., Coleman, S. L., Smith, S. K., Buxbaum, J. D., Haroutunian, V., y O'Donovan, M. C. (2004) A high proportion of polymorphisms in the promoters of brain expressed genes influences transcriptional activity. *Biochimica et biophysica acta* **1690**, 238-249.

13. Healy, J., Belanger, H., Beaulieu, P., Lariviere, M., Labuda, D., y Sinnett, D. (2007) Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood* **109**, 683-692.

14. Gerasimova, T. I., y Corces, V. G. (2001) Chromatin insulators y boundaries: effects on transcription and nuclear organization. *Annual review of genetics* **35**, 193-208.

15. Warner, J. B., Philippakis, A. A., Jaeger, S. A., He, F. S., Lin, J., y Bulyk, M. L. (2008) Systematic identification of mammalian regulatory motifs' target genes and functions. *Nature methods* **5**, 347-353.

16. Wang, X., Tomso, D. J., Chorley, B. N., Cho, H. Y., Cheung, V. G., Kleeberger, S. R., y Bell, D. A. (2007) Identification of polymorphic antioxidant response elements in the human genome. *Human molecular genetics* **16**, 1188-1200.

17. Horvath, M. M., Wang, X., Resnick, M. A., y Bell, D. A. (2007) Divergent evolution of human p53 binding sites: cell cycle versus apoptosis. *PLoS genetics* **3**, e127.

18. Echeverri, C. J., y Perrimon, N. (2006) High-throughput RNAi screening in cultured cells: a user's guide. *Nature reviews* **7**, 373-384.

19. Buckland, P. R. (2006) The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et biophysica acta* **1762**, 17-28.

20. Heyduk, T., y Heyduk, E. (2002) Molecular beacons for detecting DNA binding proteins. *Nature biotechnology* **20**, 171-176.

21. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., y Bulyk, M. L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics* **36**, 1331-1339.

22. Hudson, M. E., y Snyder, M. (2006) High-throughput methods of regulatory element discovery. *BioTechniques* **41**, 673, 675, 677 passim.

23. Gorman, C. M., Moffat, L. F., y Howard, B. H. (1982) Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells. *Molecular and cellular biology* **2**, 1044-1051.

24. de Wet, J. R., Wood, K. V., DeLuca, M., Helinski, D. R., y Subramani, S. (1987) Firefly luciferase gene: structure and expression in mammalian cells. *Molecular and cellular biology* **7**, 725-737.

25. Graham, F. L., y van der Eb, A. J. (1973) A new technique for the assay of infectivity of human adenovirus 5 DNA. *Virology* **52**, 456-467.

26. Zimmermann, U. (1982) Electric field-mediated fusion and related electrical phenomena. *Biochimica et biophysica acta* **694**, 227-277.

27. Felgner, P. L., Gadek, T. R., Holm, M., Roman, R., Chan, H. W., Wenz, M., Northrop, J. P., Ringold, G. M., y Danielsen, M. (1987) Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 7413-7417.

28. Kuppuswamy, M. N., Hoffmann, J. W., Kasper, C. K., Spitzer, S. G., Groce, S. L., y Bajaj, S. P. (1991) Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 1143-1147.

29. Lo, H. S., Wang, Z., Hu, Y., Yang, H. H., Gere, S., Buetow, K. H., y Lee, M. P. (2003) Allelic variation in gene expression is common in the human genome. *Genome research* **13**, 1855-1862.

30. Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., y Cheung, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747.

31. Knight, J. C., Keating, B. J., Rockett, K. A., y Kwiatkowski, D. P. (2003) In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nature genetics* **33**, 469-475.

32. Mottagui-Tabar, S., Faghihi, M. A., Mizuno, Y., Engstrom, P. G., Lenhard, B., Wasserman, W. W., y Wahlestedt, C. (2005) Identification of functional SNPs in the 5-prime flanking sequences of human genes. *BMC genomics* **6**, 18.

33. Galetto, R., Albajar, M., Polanco, J. I., Zakin, M. M., y Rodriguez-Rey, J. C. (2001) Identification of a peroxisome-proliferator-activated-receptor response element in the apolipoprotein E gene control region. *The Biochemical journal* **357**, 521-527.