José M. Ortiz Melón

Editor científico
e-mail: edicion@ranf.com

El Proyecto ENCODE y su trascendencia biomédica. Adaptado de Science, septiembre 7 (2012)

Cuando hace algunos años los investigadores secuenciaron el primer genoma humano, resultó una sorpresa comprobar, en primer lugar, que éste contenía muchos menos genes (codificantes de proteínas) de los esperados, en relación con su tamaño total. En segundo lugar, se encontró, que estos genes, se encontraban distribuidos (no concentrados) a lo largo de los 3000 millones de bases que constituye el DNA del genoma humano . Así pues, en lugar de los esperados 100.000 genes o más, se encontraron que solo había unos 35.000 e incluso este número se redujo posteriormente hasta unos 21.000. Entre ellos, se encontraban por tanto, millones de bases sin función aparente, y a este DNA sin función se le consideró como "DNA basura".

Recientemente, la publicación en las revistas *Science y Nature* de los resultados iniciales de un proyecto que lleva ya 10 años de trabajos denominado ENCODE(Enciclopedia de Elementos de DNA) ha venido a revelar, que contrariamente a lo que se había pensado inicialmente, cerca de un 80% del genoma humano tienen una función importante, aunque no codifique proteínas.

Además de definir mejor las regiones codificantes de proteínas, las nuevas bases funcionales reveladas por ENCODE especifican sitios de interacción de proteínas con el genoma que influencian la actividad de los genes, tiras de RNA con miles de funciones desconocidas, o simplemente lugares en los que modificaciones químicas sirven para silenciar fragmentos de nuestros cromosomas .

Tal vez lo mas importante de los datos proporcionados por el programa ENCODE, es que van a servir para clarificar factores de riesgo genéticos para una gran variedad de enfermedades y el ofrecer una mejor comprensión de la regulación y función de los genes.

En conjunto, el proyecto ENCODE ha revelado que la regulación genética es mas compleja que lo que se había pensado inicialmente, y que está influenciada

por múltiples secuencias de DNA regulador, que se localizan unas veces cerca y otras lejos de cada gen, asi como por cadenas de RNA no traducidas.

Durante los años 1990 varios investigadores habían llamado ya la atención, sobre la idea, de que el llamado "DNA basura" no era tal. Con la secuencia del genoma humano ya establecida, el Instituto de Investigación sobre el Genoma Humano de Bethesda (Maryland) decidió hace años investigar sobre cuanto de nuestro genoma, era en realidad basura y por tanto sin función. En el año 2003, se comenzó así el proyecto ENCODE en el que 35 grupos de investigación conectados entre si, proyectaron analizar 44 regiones del genoma (30 millones de bases) como proyecto piloto. Esto significaba aproximadamente un 1% de todo el genoma. En el año 2007, el proyecto piloto reveló ya que mucha de esta secuencia de DNA era activa de alguna manera, y la cuestión que se planteó entonces era conocer si el resto del genoma se comportaba como ese 1%.

Desde entonces, grupos pertenecientes a 32 instituciones de investigación repartidas por todo el mundo, han generado miles de conjuntos de datos. Mientras que los estudios piloto se llevaron a cabo por medio de una técnica llamada de microarrays basada en la comparación para analizar muestras de DNA, la fase mas amplia de la investigación ahora revelada se ha beneficiado de las nuevas tecnologías de secuenciación, que se han abaratado mucho, y han permitido avanzar mas rápidamente en este proceso .

Debido a que las partes del genoma a estudiar difieren en diferentes tipos celulares, el proyecto necesitaba ser capaz de estudiar la función del DNA en muchos tipos de células y tejidos. Al principio, el objetivo se centró en el estudio del genoma de solo tres tipos de células. Una de ellas, es una línea de leucocitos inmadura, llamada GM12878 que se ha utilizado también en un proyecto paralelo llamado el "proyecto de los 1000 genomas" que tenía por objeto caracterizar las variaciones genéticas entre humanos. En segundo lugar, la célula K562 de leucemia, y en tercer lugar la célula madre embrionaria, h1-ESC.

A medida que el proyecto se iba desarrollando los costes de secuenciación se fueron abaratando de tal manera, que se hizo posible el incorporar la secuenciación de nuevas líneas celulares. Se añadió así la línea celular de cáncer hepático HepG2 y la famosa y tradicional línea cancerosa de laboratorio, HeLaS3, así como tejido de cordón umbilical. Finalmente, otros 140 tipos de células se estudiaron también aunque en menor grado de detalle .

En todas estas células los investigadores han examinado qué bases del DNA se transcriben en RNA y si estas cadenas de RNA se transcriben en proteínas, verificando los genes codificantes de proteínas ya conocidos previamente, y localizando con mayor precisión el comienzo, el final y la región codificante de cada gen. Esta nueva reconsideración de la capacidad codificante del genoma

humano arroja ahora la cifra de que 3% del genoma contiene genes que codifica proteínas. Otros 11.224 fragmentos de DNA se clasifican como "seudogenes", es decir, genes que han funcionado como tales en algún momento de la evolución, pero que ya no están activos y sin embargo lo están en algunos tipos celulares o en algunos individuos.

El proyecto ENCODE ha revelado también el resultado de que hay muchos otros "genes" en los que el DNA codifica RNA pero este no se traduce en proteína como producto final. La gran sorpresa, es que 93% de las bases estudiadas son transcritas como RNA. En el genoma total, aproximadamente un 76% del mismo es transcrito en RNA. De todos estos RNA transcritos, 8800 tránscritos, se conocen como "pequeños RNAs" y 9900 como" largos RNAs" no codificante, y cada uno de ellos contiene aproximadamente 200 bases. Se ha llegado a conocer también, que varios de estos RNAs, se localizan en la célula en lugares determinados. Así, algunos se encuentran en el núcleo, otros en el nucleolo, otros en el citoplasma etc. A consecuencia de todo ello, algunos investigadores proponen que la unidad fundamental del genoma y por tanto la unidad básica de la herencia debe ser "el tránscrito",es decir, los fragmentos de RNA y no el "gen".

Otra manera de probar la funcionalidad del DNA es evaluar si secuencias de bases especificas están conservadas o no entre especies. Estudios previos habían mostrado que 5% del genoma humano esta conservado entre mamíferos. Algunas secuencias no conservadas entre humanos y otros mamíferos se han encontrado conservadas entre mucha gente, indicando, que un 4% adicional del genoma esta bajo selección, nuevamente, en el linaje humano, y algunas de estas regiones han podido ser relacionados con trazos distintivos de la especie humana.

Además de intervenir en el proceso de la transcripción, las bases del DNA funcionan en regulación genética a través de interacciones con factores de transcripción y otras proteínas. Diversos subprogramas han permitido localizar 3,9 millones de regiones donde los factores de transcripción se unen al genoma.

Asimismo, se ha revelado, que las nuevas regiones funcionales descubiertas solapan con bases específicas del DNA asociadas con mayor o menor frecuencia con enfermedades. Este trabajo demuestra pues, que se puede usar los datos de ENCODE para establecer nuevas hipótesis sobre la conexión entre genética y enfermedades. Así el laboratorio del Dr. Stamatoyananapoulus por ejemplo ha conseguido establecer una relación entre regiones reguladoras y sus genes específicos, localizando las variaciones en estas regiones reguladoras que incrementan riesgo de padecer una u otra enfermedad. Por ejemplo, el análisis apunta a dos tipos de células T como patogénicas en la enfermedad de Crohn, y el hecho, de que ambas están implicadas este trastorno inflamatorio del intestino.

Es de esperar que este tipo de correlaciones se pueda ir extendiendo a otras muchas enfermedades genéticamente complejas y con ello poder ir definiendo mejor los llamados factores de riesgo.