# Un año de ENCODE

El Proyecto ENCODE (*ENCyclopedia Of DNA Elements*), financiado por el NHGRI (*National Human Genome Research Institute, USA*), se marcó como objetivo la identificación de todas las regiones de transcripción, de asociación a factores de transcripción, estructuras de cromatina y modificaciones de histonas en la secuencia del genoma humano. Gracias a la identificación de estos elementos funcionales, actualmente el 80% de los componentes del genoma humano tienen ya al menos asociada una función bioquímica. El 5 de Septiembre de 2012, el grueso de los resultados de ENCODE se hizo libre y accesible a todo el mundo a través de la aplicación *Nature ENCODE Explorer,* accesible en: http://www.nature.com/encode/

Hace unos meses pedimos a tres estudiantes de Biología que escribieran un comentario sobre el Proyecto ENCODE. Ahora, celebrando el aniversario del lanzamiento de *Nature ENCODE Explorer*, *Encuentros en la Biología* lo publica.



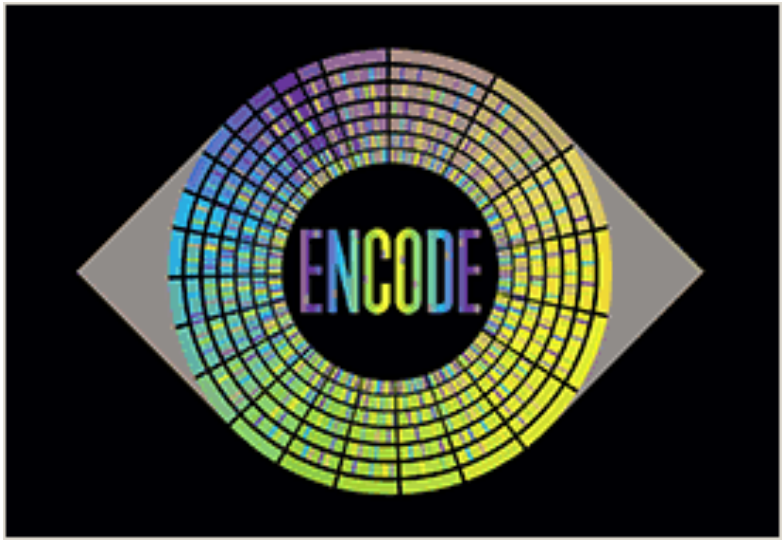



Imagen que acompañó el lanzamiento del *Nature* ENCODE explorer el 5 de Septiembre de 2012

## Proyecto ENCODE

José Joaquín Serrano Morales, Mª Carmen Ocaña Farfán, Elena Díaz Santiago
Alumnos de último curso de Biología, Universidad de Málaga
conguino@hotmail.com
meru_wolf@hotmail.com elenads@msn.com

La entrada de siglo nos trajo uno de los momentos más esperados por la comunidad científica: la secuenciación del genoma humano. En 1990 comenzó el llamado Proyecto del Genoma Humano, que logró secuenciar en 2003 el primer genoma humano completo. La expectación radicaba en los prometedores logros que se llevarían a cabo tras semejante avance en el conocimiento. Tras este gran evento para la ciencia, comenzó el denominado Proyecto de los 1000 Genomas, que finalizó en octubre de 2012 con la secuenciación de 1092 genomas de individuos diferentes. Cuál fue la sorpresa para todos cuando se percataron de que en realidad no habían hecho nada más que rasgar la punta del iceberg. No solo la solución no estaba en los genes, sino que el compendio de mecanismos que llevan asociados para regular su expresión se presentaba casi inexpugnable. Pero la comunidad científica se mueve con retos y el reto se ha convertido en hazaña con la publicación en el año 2012 (más de una década después de la secuenciación del genoma humano) del proyecto ENCODE, lanzado por el NHGRI (*National Human Genome Research Institute*) de Estados Unidos. ENCODE es el acrónimo en inglés de *Encyclopedy of DNA Elements*, cuyo origen se remonta al año 2003. En su fase inicial ENCODE abarcó tan solo un 1% del genoma; ya en el año 2007 se publican los primeros resultados. El proyecto culmina con el análisis completo del genoma cinco años después. Este proyecto se ha podido llevar a cabo gracias a las nuevas tecnologías de secuenciación que permiten secuenciar genomas enteros en tiempos impensables

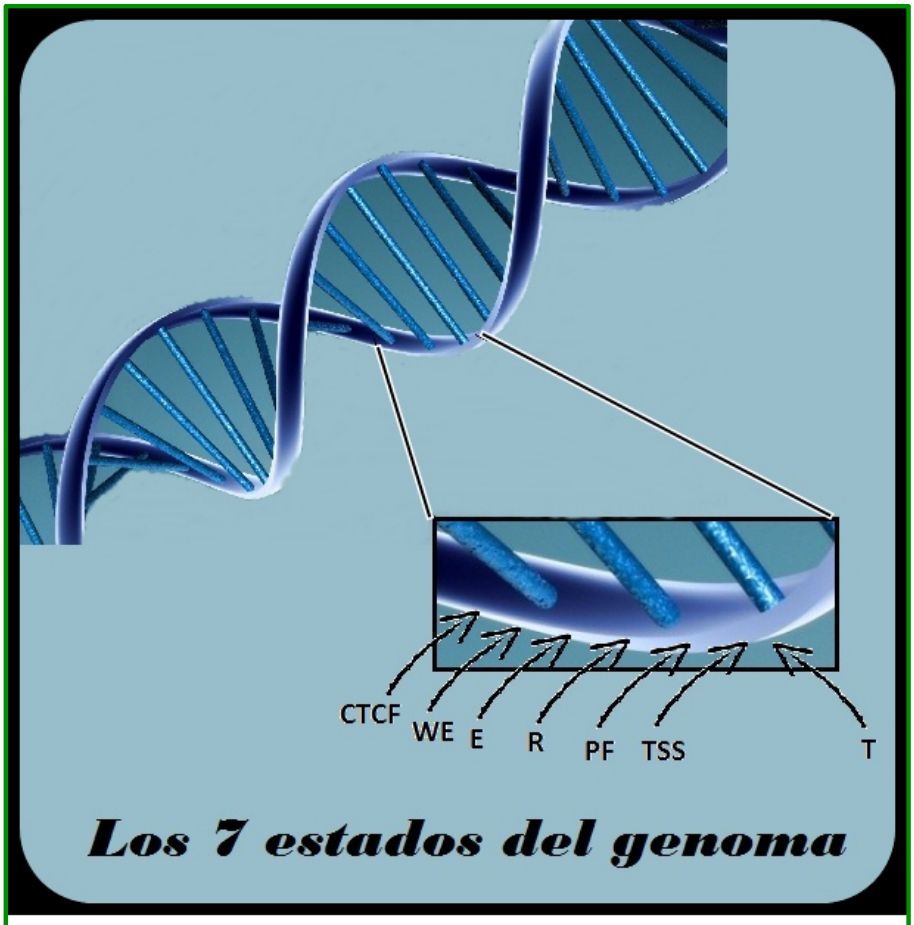
hace solo una década además de realizar análisis más precisos, con un amplio repertorio de ensayos funcionales. La automatización que brinda la gran cantidad de *software* informático para realizar estudios comparativos ha permitido que un total de 442 investigadores pertenecientes a 32 instituciones distintas hayan realizado y estudiado la actividad del genoma completo de 147 tipos celulares distintos y que sus resultados fueran publicados por separado en la revista *Genome Research*. Al mismo tiempo la revista *Nature* ha publicado una visión general del proyecto, en septiembre del año 2012. Estos artículos y los datos generados son de libre acceso para cualquier científico de cualquier lugar del mundo. Para todos aquellos interesados en acceder en más profundidad aquí dejamos el enlace de la página desde la cual se puede acceder a los distintos artículos:

http://www.nature.com/encode/#/.

¿Pero cuál es la intención del proyecto? ENCODE pretende localizar en el genoma los elementos funcionales del mismo. Un elemento funcional es todo aquel segmento discreto del genoma que codifica un producto definido o que muestra una función bioquímica determinada. Entre estos elementos funcionales se encuentran regiones transcritas de ARN, regiones codificantes para proteínas, sitios de unión a factores de transcripción, estructura de la cromatina y modificación de histonas, con lo cual no se incluyen únicamente genes, sino también elementos reguladores, ARNs no codificantes, transcritos con *splicing* o ayuste alternativo, etc. Todos estos elementos nos ponen sobre aviso de la poca importancia que hasta ahora se le había dado al mal llamado ADN "basura". Sin embargo, gracias al proyecto ENCODE sabemos que el 80% del genoma tiene una función, es decir, tiene una actividad bioquímica específica. Formando parte del proyecto ENCODE encontramos a GENCODE, que se encarga del estudio de los loci codificantes y transcritos con *splicing* alternativo, *loci* no codificantes con evidencias de transcripción, y pseudogenes.

Analizando 1640 bases de datos e integrando los resultados de los experimentos llevados a cabo en los 147 tipos celulares diferentes por los distintos laboratorios y asociaciones se han descrito importantes características acerca de la organización y la función del genoma humano. Quizá la más relevante sea la división del genoma en siete estados de la cromatina: CTCF (del inglés, *CCCTC-binding Factor*), E (*Enhacer*), PF (*Promoter Flanking region*), R (*Repressed region*), TSS (*Transcription Start Site*), T (*Transcribed region*) y WE (*Weak Enhacer*). Tres de ellos son estados "activos", en concreto, los estados E y WE, que son regiones potenciadoras que se diferencian en la fuerza de la potenciación, y el estado CTCF, que posee numerosos sitios






Los 7 estados del genoma: CTCF (CCCTC-binding factor), WE (weak enhancer), E (enhancer), R (repressed region), PF (promoter flanking region), TSS (transcription start site) y T (transcribed region). El orden en el que se disponen en la figura no es necesariamente su orden en la realidad, pero podría ser una disposición lógica.

de unión a CTCF, encargado de la regulación de la estructura de la cromatina. Por otro lado, hay un estado represor, el estado R; el estado TSS corresponde a los promotores, incluyendo los sitios de inicio de la transcripción, y enriquecido en factores de transcripción que actúan cerca de los promotores y en las ADN polimerasas II y III; un estado PF, que incluye regiones flanqueantes a los promotores y un estado T, que constituye las regiones transcritas. La diferencia entre unos estados del genoma y otros se basó en la observación de los patrones de accesibilidad a la cromatina, y de diferentes modificaciones de las histonas en residuos de lisina específicos, que conllevan la asociación de diferentes marcadores de acetilación y metilación, como pueden ser H3K27ac o H3K4me1. La acetilación y la metilación poseen funciones antagónicas, de forma que la acetilación favorece la transcripción porque relaja la cromatina, mientras que la metilación reprime la transcripción. Así, se han visto diferentes niveles de metilación en los diferentes estados del genoma, presentando una mayor metilación los estados que forman parte de genes, es decir, el estado T, y un menor grado de metilación en las regiones promotoras, como TSS. Además, hay diferencias de metilación entre los dos estados potenciadores, estando más metilado el estado WE, menos activo. Con ello, puede deducirse que la metilación en los promotores da lugar a su represión, mientras que en genes aumenta la actividad transcripcional. Es obvio pues que el estado WE presente un mayor grado de metilación que el estado E, ya que es menos activo.

Por otra parte, las regiones de lo que se llama cromatina abierta están caracterizadas por presentar hipersensibilidad a la DNAasa I, que además es característico de regiones reguladoras, por lo que es lógico pensar que, por ejemplo, el estado CTCF represente zonas de cromatina abierta.

Otras aportaciones interesantes de este proyecto son la afirmación de que la modificación de histonas y la unión a factores de transcripción están implicados de alguna forma en la transcripción y, por tanto, en el nivel de expresión, y además que hay suficiente información en los promotores como para explicar la mayoría de la variación en la expresión del ARN. Se trata también la distribución de los sitios de unión a factores de transcripción, que no se distribuirían de manera aleatoria a lo largo del genoma, sino que lo hacen en función de los promotores y de otros sitios de unión a factores de transcripción.

Del proyecto ENCODE también podemos destacar sus estudios acerca de la variación genómica humana, analizando diferentes variantes en los elementos funcionales tratados por ENCODE. Para ello, utilizaron el genoma de un individuo que fue secuenciado en el proyecto de los 1000 genomas, además del genoma de sus padres, de forma que pudiera comprobarse la variación alelo-específica. Con esto, se comprobó que un 1% de los sitios de unión a factores de transcripción son específicos del haplotipo de uno de los parentales, de forma que no se encuentra presente en el otro.

La última aplicación importante que mencionaremos del proyecto ENCODE es la relación entre variantes genómicas asociadas a enfermedad. Para su estudio, se analizaron diferentes polimorfismos de nucleótido simple (SNPs) que dan lugar a un fenotipo concreto, de los cuales la gran mayoría son regiones intrónicas o intergénicas, por lo que podemos deducir que una enfermedad no tiene que venir dada de forma obligada por un gen o una modificación de éste. En estos estudios se vio que una gran proporción de SNPs solapan con sitios hipersensibles a DNAasa I, con lo que se deduce que estos SNPs forman parte de regiones reguladoras, y que también coinciden con sitios de unión a factores de transcripción. Es interesante el caso de un *gene desert* (grandes fragmentos de ADN que carecen de regiones codificantes para proteínas) en el cromosoma 5 que contiene ocho SNPs asociados a enfermedades inflamatorias, como la enfermedad de Crohn. En concreto, los datos recogidos por ENCODE refuerzan la hipótesis de que las variantes genéticas de este fragmento del cromosoma modulan la expresión de los genes flanqueantes y que, además, estas variantes afectan de manera alelo-específica al nivel de ocupación del factor de transcripción GATA, lo cual influye en la susceptibilidad a la enfermedad de Crohn. Así, estos SNPs influirían en las variantes funcionales y, por tanto, llegarían a poseer una función de forma indirecta.

El proyecto ENCODE provee de una fuente importante de datos para la comunidad científica, y además ha permitido un mayor entendimiento del genoma humano gracias al análisis de múltiples elementos funcionales, descubriéndose varios aspectos acerca de la expresión y la regulación. Sin embargo, los mismos autores admiten que los resultados no pueden tomarse en cuenta en las proporciones que se dan, ya que los experimentos se realizan sobre dos líneas celulares, de forma que realmente los datos deberían infravalorarse. En tan solo cinco meses ENCODE ha recibido 52 citaciones en *Web of science*, 108 en *CrossRef* y 73 en *Scopus*, algunas de las principales bases de datos de la comunidad científica, lo que nos indica la enorme repercusión que está teniendo la publicación de estos datos dentro de la comunidad científica.

Como estudiantes de Biología tenemos que decir a todo aquel que pretenda profundizar en el tema que los artículos que recogen toda la información referente a ENCODE presentan mucha información en forma de gráficas complejas y tablas de datos, lo cual hace que estos artículos sean difíciles de leer y comprender para un estudiante, estando más orientado, por tanto, a especialistas en la materia y personal docente e investigador en el ámbito de la Genética y la Biología Molecular.



**NOTA DE LOS EDITORES:** La publicación de los resultados del Proyecto ENCODE ha suscitado un vivo debate (del que dan cuenta algunas de las más destacas revistas de investigación y algunos de los más frecuentados foros de la ciencia) acerca de las fortalezas y debilidades de los proyectos a gran escala con inversión cien/mil millonaria en el actual contexto de la ciencia.