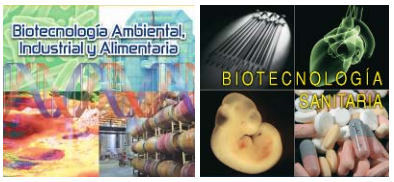
Revisión

El proyecto *ENCODE* desvela la complejidad del genoma humano



Beyond Darwin*
Universidad Pablo de Olavide de Sevilla

Palabras clave: proyecto ENCODE; genoma humano; epigenética

RESUMEN

A pesar de conocer la secuencia completa del genoma humano desde 2001, el conocimiento que se tiene de nuestro genoma es aún bastante limitado, sobre todo, en el terreno de los RNAs no codificantes, los transcritos que sufren *splicing* alternativo las secuencias reguladoras así como sobre la organización tridimensional del mismo. En el proyecto *ENCODE* publicado recientemente en la revista Nature y que revisamos en este articulo, más de 200 científicos de todo el mundo han participado de forma coordinada para desvelar los secretos de nuestro genoma (Ecker et al., 2012).
El proyecto *ENCODE* ha cambiado la forma de ver nuestro propio genoma, dejando atrás el viejo concepto de *DNA basura* con el que se conocía lo que desconocíamos. ENCODE ha puesto de manifiesto el nivel complejidad del genoma humano, sus elementos reguladores y su intrincada interconexión tridimensional. El conocimiento arrojado por este proyecto tendrá una importante repercusión sobre el diagnóstico y la terapia de numerosas enfermedades humanas.

1. RESULTADOS

1.1 Identificación de elementos funcionales y el transcriptoma

El objetivo principal del proyecto *ENCODE* es aclarar el conjunto de elementos funcionales codificados por el genoma humano. Con este término no solo se alude a los segmentos del genoma que típicamente codifican productos específicos (como proteínas y RNAs no codificantes), sino también a aquellos que muestran propiedades bioquímicas características, como, por ejemplo, la capacidad de unión a determinadas proteínas o las estructuras específicas que adquiere la cromatina (Ecker et al., 2012).

1.1.1 Detección de transcritos registrados en GENCODE *y nuevos transcritos*

La mayoría de los genes que ya se encontraban registrados en *GENCODE*, un sub-proyecto de *ENCODE* de anotación génica, se hallaron en las muestras génicas previamente investigadas. Los nuevos transcritos descubiertos cubren en total el 78% de los nucleótidos intrónicos, así como un 34% de las secuencias intergénicas. Esto implica un aumento del número de regiones intergénicas con respecto a las que se conocían, además de incrementar la cantidad de exones, sitios de *splicing*, transcritos y genes de *GENCODE* preexistentes. Sin embargo, el incremento del número de genes identificados es mucho mayor que el de lugares de *splicing*, debido mayoritariamente a la detección de transcritos monoexónicos. Se piensa que esto podría estar causado por una leve contaminación por ADN o bien por una determinación incompleta de los transcritos.

1.1.2 El transcriptoma de subcompartimentos nucleares y la localización subcelular final de los distintos RNAs

En una determinada línea celular se analizó el ARN de tres subcompartimentos nucleares: cromatina, nucleolo y nucleoplasma. Casi la mitad de los genes se detectaron en estos tres subcompartimentos nucleares, y sólo una pequeña fracción resultó ser específica de un determinado subcompartimento. Al analizar el ARN en diferentes compartimentos subcelulares, se observó que hay un orden de magnitud más de expresión génica en ARNs poliadenilados respecto a los que no lo están. La expresión génica también es generalmente mayor para aquellos genes que codifican proteínas en comparación con los que producen ARNs no codificantes (lncRNAs), aunque no obstante, en determinadas líneas celulares ambos tipos exhiben el mismo nivel de expresión (Djebali et al., 2012).

Cada tipo de ARN se localiza en los compartimentos celulares en los que cumple sus correspondientes funciones. Según el estudio realizado, sólo los genes que codifican proteínas parecen distribuirse preferencialmente en el citosol, mientras que en el núcleo se sitúan principalmente lncRNAs. El resto, a excepción de ciertos tipos, como pseudogenes y algunos ncRNAs pequeños, se distribuyen de forma más o menos uniforme entre ambos compartimentos. La expresión de lncRNAs comunes a todas las células es menor que la de ARNs codificantes, mientras que la expresión de los lncRNAs específicos de línea celular sí es mayor que la de codificantes. También, se han encontrado proporciones significativas de transcritos codificantes y no codificantes procesados a RNAs estables de menos de 200 nucleótidos (tRNA, miRNA, snRNA y snoRNA). miRNAs y tRNAs en el citosol, snoRNAs en el núcleo y snRNAs en ambos compartimentos, aunque de forma específica en la cromatina cuando aparece en el núcleo, coincidiendo con el hecho de que el proceso de *splicing* sea predominantemente cotranscripcional.

1.1.3 Catálogo de RNAs del genoma humano

Mediante el análisis del estado de procesamiento de las diferentes fracciones de ARN, se confirmó que el *splicing* ocurre predominantemente durante la transcripción. Esto apoya las pruebas que relacionan la estructura de la cromatina con la regulación del *splicing*, ya que la cromatina que codifica los transcritos que se procesan posee un mayor número de marcas de regulación.

El proyecto ha conseguido secuenciar RNAs de diferentes líneas celulares con el fin de desarrollar un extenso catálogo de su expresión. De las secuencias de ARN nuevas, una fracción coincidía con subfragmentos de otros longRNAs ya conocidos, especialmente tRNAs. Este fenómeno ya había sido descrito en G*ENCODE*, e indica que probablemente sean precursores. Se comprobó más específicamente que, en comparación, los exones de lncRNAs suelen contener mayor cantidad de snoRNAs. El resto de secuencias se consideraron nuevas, perteneciendo la mayoría a una única línea celular cada una, y estando casi el 40% de estos nuevos RNAs asociados con regiones promotoras o terminadoras de genes ya descritos previamente.

La mayoría de longRNAs, tanto nuevos como ya descubiertos con anterioridad, parecen tener papeles funcionales importantes, tanto en la codificación de proteínas como en actuar de precursores para smallRNA. Para comprobarlo se comparó su localización en una línea celular concreta. En el caso de miRNAs, tRNAs y snRNAs, los precursores correspondientes resultaron ser predominantemente nucleares. En cambio, los longRNAs asociados a snoRNAs aparecían en ambos compartimentos (Djebali et al., 2012; Neph et al., 2012).

1.1.4 Iniciación y terminación de la transcripción

Para el estudio del inicio y la terminación de la transcripción se empleó el método *CAGE* (*cap-analysis of gene expression*), que permite analizar las zonas de inicio de la transcripción (TSS o *transcription start site*), así como sus promotores asociados a través del reconocimiento de señales en el extremo 5'. De esta manera, se han identificado una enorme cantidad de sitios de inicio de la transcripción (Djebali et al., 2012; Neph et al., 2012).

Analizando ARN poliadenilado, se detectó una mayoría de TSS registrados en G*ENCODE*, y aproximadamente un 20% de TSS nuevos: intergénicos o antisentido. Un 72% de las secuencias analizadas por *CAGE* eran TSS, por lo que el resto debían de ser fruto del "recaperuzado" o bien de un nuevo tipo de TSS. Mediante la comparación entre los TSS registrados en G*ENCODE* y los determinados por *CAGE*, se relacionaron ambos con aspectos característicos del inicio de la transcripción, como son la hipersensibilidad a DNAasa, la remodelación de la cromatina o la presencia de elementos unidos a DNA. Un 45% de los TSS registrados en G*ENCODE* también respondieron al *CAGE*, y aproximadamente la mitad de éstos resultaron estar asociados con al menos una de estas características, de forma que sólo una pequeña parte presentaba todas estas características simultáneamente. Esto sugiere la posibilidad de que las regiones reguladoras próximas a los TSS sean de más de un tipo.

Por otra parte, analizando el extremo 3' se identificaron alrededor de 128.000 sitios potenciales de poliadenilación dentro de los transcritos registrados en G*ENCODE*. Alrededor del 20% eran sitios ya registrados, mientras que el 80% correspondían a nuevos PAS (*poliadenylation sites*), lo cual supone un aumento considerable.

1.2 Elementos reguladores

Con el proyecto *ENCODE* se han conseguido elucidar importantes características de la organización y función del genoma humano. Una de las ideas más importantes es que un 80'4% del genoma humano (la gran mayoría) participa en al menos un proceso regulador, pero diferencialmente, según el tipo celular. Asimismo, la mayor parte del genoma se sitúa cerca de algún elemento regulador. De hecho, el 95% del genoma se encuentra a un máximo de ocho kilobases de un sitio de interacción DNA-proteína y un 99% se encuentra a un máximo de 1'7 kilobases de al menos uno de los elementos bioquímicos detectados en el proyecto. En otras palabras, la regulación de la expresión se encuentra prácticamente extendida por todo el genoma y lo que antes se denominaba DNA basura se ha revelado como un elemento esencial en la regulación de la genética de los seres vivos (Dunham et al., 2012).

1.2.1 Identificación de enhancers *y promotores*

El proyecto *ENCODE* ha permitido identificar un total de 399.124 regiones que actúan como *enhancers*, otras 70.292 que tienen características de promotores y cientos de miles de zonas quiescentes. Igualmente, ha permitido correlacionar cuantitativamente tanto la producción y el procesamiento de RNAs con marcas situadas en la cromatina, como la unión de factores de transcripción a promotores, con resultados que indican que la funcionalidad de los promotores puede explicar la mayor parte de las variaciones en la expresión del RNA.

1.2.2 La importancia de los elementos funcionales no codificantes

Otra idea interesante que se desprende de este trabajo es que gran parte de la variabilidad no codificante en genomas individuales, es decir, las variantes de algunas regiones que no tienen efectos en el transcrito, se encuentran en regiones funcionales, mientras que los polimorfismos de un solo nucleótido asociados a enfermedades se encuentran en mayor medida en elementos funcionales no codificantes. Gracias a *ENCODE*, se ha permitido conocer la existencia de polimorfismos en un solo nucleótido, los cuales se localizan generalmente en regiones cercanas a *enhancers* y promotores. De este modo, se ha podido determinar que algunos de estos polimorfismos se producen en regiones de unión a factores de transcripción, provocando así enfermedades tales como la enfermedad de Crohn.

1.2.3 Las regiones codificantes y el splicing

Por otro lado, se han identificado una gran cantidad de regiones codificantes de proteínas y de RNAs no codificantes, así como pseudogenes. Algunos de estos, curiosamente, están asociados a la cromatina activa y, por lo tanto, son transcritos.

Sorprendentemente, los investigadores han establecido que los 20.687 genes codificantes de proteínas identificados pueden dar lugar a una media de 6'3 transcritos diferentes debido al *splicing* alternativo, pudiendo generar hasta 130.000 productos diferentes. Del análisis de los patrones de *splicing* y la expresión de isoformas alternativas se concluyó que los genes tienden a expresar varias isoformas simultáneamente, generalmente hasta 10 o 12 isoformas por gen. Sin embargo, la metodología empleada no ha permitido distinguir si la expresión de estas isoformas se debe a su producción en una misma célula, o de forma específica en cada tipo celular. No todas las isoformas están representadas uniformemente, sino que generalmente una domina frente a las demás. Alrededor de tres cuartas partes de los genes que codifican proteínas tienen al menos dos isoformas dominantes, en una determinada línea celular.

Además, en algunas de las líneas celulares se han identificado secuencias peptídicas en regiones intergénicas, lo cual indica que todavía quedan genes codificantes de proteínas por encontrarse.

1.3 Huellas en el ADN

1.3.1 La accesibilidad al ADN y los sitios hipersensibles a DNasaI (DHSs)

En este estudio, también se ha caracterizado la accesibilidad a distintas zonas de la cromatina mediante la hipersensibilidad a cortes producidos por la DNasaI, la cual también puede actuar como indicador de las regiones reguladoras del DNA (Neph et al., 2012). Así, se ha generado el primer mapa de los DHSs humanos, tras el estudio de 125 tipos de celulares diferentes, tales como células primarias normales diferenciadas, células inmortalizadas, líneas celulares derivadas de tumores malignos y células progenitoras multipotentes y pluripotentes. Se identificaron 2.890.742 DHSs distintos, de los cuales 970.100 eran específicos

de un solo tipo celular, mientras que 1.920.642 estaban en dos o más tipos de células. Aproximadamente, el 5% de las huellas dejadas en el ADN se localizó en lugares de inicio de la transcripción, encontrándose el 95% restante en zonas más dístales y repartidos más o menos de una forma equitativa entre zonas intra e intergénicas (Neph et al., 2012).

Además, se han empleado estos ensayos de hipersensibilidad en diferentes tipos celulares para la identificación de 8'4 millones de huellas dejadas por diversas proteínas sobre el DNA. De estas huellas, se han recuperado un 90% de motivos de factores de transcripción conocidos y cientos de otros nuevos. Esta información va a permitir el estudio nuevas relaciones entre la accesibilidad de la cromatina, la transcripción, la metilación y los patrones de ocupación de los elementos reguladores (Janicki et al., 2004).

Para estudiar cómo la evolución ha ido dando forma al conjunto de zonas DHS en los seres humanos, se estudió la diversidad de nucleótidos de las mismas en 53 individuos no relacionados. Para obtener una comparación con sitios no sometidos a selección natural (aquellos en los que las variaciones de ADN se clasifican como neutras), se comparó con la diversidad de nucleótidos de los sitios cuatro veces degenerados (aquellos en los que todos los cambios posibles son sinónimos). Se concluyó que los sitios DHSs mostraban una diversidad menor que los sitios cuatro veces degenerados, lo que concuerda con una selección negativa o purificadora, un tipo de selección natural en el que la diversidad genética decae según un valor particular del carácter.

Otro punto que llamó la atención de los investigadores del proyecto *ENCODE* fue que la diversidad de nucleótidos de las zonas DHS aumentaba con la capacidad proliferativa de las células. Para explicar el origen de este fenómeno realizaron experimentos con células humanas con diferente capacidad proliferativa y compararon los resultados con experimentos similares realizados en chimpancés, llegando a la conclusión de que la tasa de mutación relativa aumenta con la capacidad proliferativa de la célula, lo que descubre un vínculo insospechado entre la accesibilidad a la cromatina, el potencial proliferativo y la variación humana.

1.3.2 Interacción entre factores de transcripción y el ADN

Otro de los aspectos investigados en el proyecto *ENCODE* ha sido la determinación de los sitios de unión de los factores de transcripción al ADN. La unión de factores de transcripción al ADN puede ser detectada como una huella o "footprint" en la propia secuencia del ADN cuando éste es tratado con DNasaI. El footprint se produce como consecuencia de una actividad desigual por parte de la enzima a lo largo del ADN no unido a histonas, el cual es degradado por la enzima, excepto en aquellas zonas en las que se encuentra unido a diferentes proteínas (Neph et al., 2012).

La unión de estos factores se produce a lo largo de todo el genoma, ocupando tipos de ADN que difieren en función en un porcentaje variable: promotores, UTR, intrones, regiones intergénicas e incluso regiones codificantes, en las cuales limita la variabilidad al determinar el tipo de codones a usar, ya que las proteínas que se unen son específicas de nucleótidos, no de codones. Además, existe una relación entre el patrón epigenético, como la metilación de las islas CpG, y los lugares que presentan footprint, ya que aquellas citosinas metiladas pierden la capacidad de unir factores de transcripción y, por lo tanto, de regular la expresión genética (Goldberg et al., 2007).

El ensayo de hipersensibilidad aplicado a estructuras proteína-ADN cristalizadas también ha servido para conocer de manera más precisa la morfología de la unión factor de transcripción-ADN, además de caracterizar el lugar de inicio de la transcripción como un footprint hallado corriente arriba de la mayoría de los genes. Por otro lado, llevando a cabo experimentos de forma paralela, como inmunoprecipitación de cromatina y posterior secuenciación de las secuencias recuperadas (*ChIP-seq*) se ha esclarecido que la interacción proteína-ADN puede ser directa, cuando la proteína posee el motivo de unión a los nucleótidos, o indirecta, si se une a una segunda proteína que lo posee. El modelo de unión difiere entre los distintos tipos celulares, determinando la diferenciación celular. Los motivos de unión entre proteínas y ADN se conservan entre humanos y otros vertebrados como consecuencia de una selección de aquellos que resultaban más estables y eficaces.

1.3.3 Predicción de expresión génica mediante el análisis las marcas epigenéticas en los promotores

Gracias al programa *ENCODE*, se han podido esclarecer muchos interrogantes sobre los promotores y su regulación. De ahí que se haya comprobado que los niveles de expresión de RNA pueden ser predichos simplemente a partir de patrones de modificaciones de la cromatina, los cuales básicamente consisten en marcas de acetilación y metilación en residuos concretos de lisina de las histonas, o de la presencia de factores de transcripción (Sanyal et al., 2012; Thurman et al., 2012).

Los factores de transcripción presentan una alta afinidad por las regiones ricas en islas CpG cercanas a los promotores, los cuales se han sido clasificados en: anchos (sin caja TATA y ricos en nucleótidos C-G) y estrechos (con caja TATA). Estos factores se pueden unir a regiones cromosómicas con diferentes estructuras y marcas en la cromatina. Los resultados indicaron que la asociación factor de transcripción-isla CpG no solo se correlaciona con la orientación de los promotores, sino que también se relaciona con patrones direccionales y estructurales de las modificaciones de las histonas. Igualmente, se confirmó que los factores transcripcionales constituyen barreras que alternan modificaciones en nucleosomas e histonas dentro de un rango de posibles configuraciones. Las regiones de unión a los factores transcripcionales presentan pautas de asociación, por lo que no se distribuyen de manera aleatoria. Se han identificado regiones llamadas HOT que presentan una alta proporción de estas secuencias ricas en sitios de unión de factores.

1.3.4 Determinación de los orígenes de transcripción

Otro aspecto abordado por el proyecto *ENCODE* ha sido la relación entre los orígenes de transcripción (TSS) y las modificaciones H3K4me3. Esto ha permitido estudiar la relación entre la accesibilidad de la cromatina y los patrones de las modificaciones H3K4me3 en los promotores caracterizados, atendiendo también a su variabilidad entre los tipos celulares estudiados (Thurman et al., 2012).

Para analizar esta relación se realizaron secuenciaciones ChIP (*chromatin immunoprecipitation*) y *high-throughput* en las regiones cromatínicas con H3K4me3, utilizando los mismos tipos celulares usados para el análisis con DNasaI. Comparando la densidad de lugares de corte de la DNasaI con los marcadores de la secuenciación ChIP de las proximidades de los TSSs, se encontró un patrón asimétrico altamente estereotipado, lo que demuestra una relación precisa con los TSSs. Este patrón direccional se debe a la presencia de nucleosomas anclados corriente abajo del promotor DHS, y presenta poca variación entre las distintas líneas celulares.

Para encontrar promotores desconocidos escanearon el genoma de 56 líneas celulares buscando coincidencias con este patrón. Se identificaron 44.853 nuevos promotores posibles que presentaban todas las configuraciones posibles, algunas incluso en antisentido respecto a la dirección de la transcripción o inmediatamente a continuación del extremo 3' de un gen. Esto sugiere la existencia de un gran grupo de promotores transcripcionales específicos de cada célula, de los cuales muchos pueden presentar orientación antisentido (Thurman et al., 2012).

1.3.5 Metilación del DNA

La relación entre la metilación del DNA y la estructura de la cromatina aún no se ha definido claramente, y este es uno de los puntos importantes de esta investigación. Por lo general, la metilación de las citosinas se produce en dinucleótidos CpG y está implicada en la regulación epigenética de la expresión génica. Frecuentemente, las islas CpG más metiladas se encuentran en los genes y las regiones intergénicas, en lugar de en los promotores y en las regiones aguas arriba de regulación. Típicamente, la metilación del promotor está asociada con la represión, mientras que la metilación génica se correlaciona con expresión génica. Gracias a la impronta genómica, se han podido identificar CpGs metiladas de alelos específicos, demostrándose así que estos loci muestran una metilación aberrante en líneas celulares cancerosas. Además, se han detectado metilaciones poco comunes de citosinas fuera de dinucleótidos CpG en tejidos adultos que pueden desempeñar importantes importantes funciones (Djebali et al., 2012; Dunham et al., 2012).

Secuenciando un cierto número de patrones de metilación, se ha confirmado que existen patrones de metilación diferenciales según el tipo celular y que los niveles de metilación del DNA están correlacionados con la accesibilidad a la cromatina. El análisis comparativo del patrón de metilación y la accesibilidad de la cromatina en 19 tipos celulares reveló la existencia de 243.057 DHSs, y permitió su clasificación en dos grupos: los que tienen

una fuerte correlación inversa entre la metilación del ADN y la accesibilidad de la cromatina en diversos tipos de células, y en los que la accesibilidad a la cromatina varía pero presentan hipometilación constitutiva. Para cuantificar estas tendencias, se realizó un análisis de regresión lineal entre la accesibilidad a la cromatina y la metilación del ADN, mostrándose que un 20% tenía una asociación significativa entre metilación y accesibilidad: por lo general, la metilación de la cromatina conlleva una limitación en su accesibilidad.

1.3.6 Interacciones génicas de corto y largo alcance

En relación a la expresión génica, otro de los aspectos abordados en el programa *ENCODE* fue la interacción física que puede darse entre distintas regiones cromosómicas, aunque estas estén separadas incluso por cientos de kilobases. Tanto es así que las regiones promotoras de la mayoría de los genes están formando parte de complejos de interacción multi-gen, los cuales pueden abarcar hasta varias megabases. Estas interacciones de largo alcance solo se pudieron observar en solo una de las cuatro líneas celulares estudiadas, de lo que se deduce el alto grado de especificidad del tejido en su manera de conectar los genes con otros elementos (Gerstein et al., 2012).

Un análisis exhaustivo de las interacciones entre distintas regiones del genoma, conocidas como *looping interactions*, o interacciones que ocurren entre promotores y elementos alejados de éstos en el genoma, ha demostrado que existen interacciones a larga distancia entre distintas regiones del genoma tienen funciones importantes en la regulación de la expresión génica. Por ejemplo, se ha visto una fuerte correlación entre la interacción de algunos pares de loci, como los TSS de los genes y ciertos *enhancers*, con la expresión de los genes que regulan. El número medio de elementos dístales que interactuaron con un TSS es de 3,9, y el número promedio de TSSs que lo hicieron con un elemento distal es de 2,5. Con estos resultados, los investigadores sospechan que ha de existir una compleja red de cromatina interconectada.

En el proyecto *ENCODE*, se analizaron 14 regiones génicas mediante la técnica 5C, lo que permitió establecer las primeras reglas de interacción genómica. Las interacciones entre elementos del genoma dentro de un mismo fragmento son más frecuentes, mientras que dentro de un mismo fragmento, las interacciones ocurren más frecuentemente entre elementos situados más próximos entre sí. Las interacciones entre elementos localizados en fragmentos dentro de un mismo cromosoma ocurren con mayor frecuencia que en fragmentos en localizaciones cromosómicas distintas. La tendencia parece clara: cuanto más próximos estén los distintos fragmentos, más frecuente serán las posibles interacciones entre ellos.

Quizás una de las conclusiones más importantes que se obtuvieron durante el desarrollo de esta parte del proyecto, fue el efecto de la especificidad celular sobre las interacciones entre las regiones TSS y otros elementos distantes. Alrededor del 60% de las interacciones detectadas fueron observadas sólo en una de las líneas celulares usadas. En este caso, se vuelve a hacer patente que la especificidad celular interviene en el proceso de plegamiento y en la disposición tridimensional del genoma en la célula.

Por otro lado, se vio que la mayor parte de las interacciones físicas usuales en todas las líneas celulares estudiadas tienen en común, en frecuencias muy altas, la interacción de los TSS con fragmentos distantes unidos a CTCFs (*CCCTC-Binding factor*), que tienen la capacidad de regular la compactación del DNA. Mediante el uso de técnicas de mapeo génico y secuenciación como FAIRE o DHS, se han observado que estas interacciones implican la "descompactación" de la cromatina (formación de la eucromatina) y la modificación de elementos como histonas (H3K4me1, H23K4me2 y H3K4me3), que indican el inicio de la actividad transcripcional de esa región del genoma.

1.3.7 Los siete estados genómicos

Con objeto de tener una visión más global de la regulación de la expresión génica, el proyecto *ENCODE* también ha identificado 13.000 *enhancers*, se han estudiado grupo de histonas modificadas y diversos patrones de acceso a la cromatina en distintas líneas celulares (Dunham et al., 2012). Todo ello ha permitido definir siete estados genómicos distintos:

- CTCF: sin modificaciones de histonas y asociado a la cromatina abierta. No varía en las distintas líneas celulares.

Algunos números del proyecto ENCODE sobre el genoma humano

Nº de genes codificantes	20.687
Nº medio de transcritos por gen (splicing alternativo)	6'3
Nº Enhancers	399.124
Nº Promotores funcionales	70.292
Nº Promotores quiescentes o crípticos	>100.000
Nº Sitios hipersensibles a DNasaI (DHSs)	2,89 millones
Nº de Huellas dejadas por diversas proteínas*	8,4 millones
% del genoma que participa en al menos un suceso regulador**	80'4
% del genoma se encuentra a un máximo de ocho kb de un sitio de interacción DNA-Proteína	95
% del genoma que se encuentra a un máximo de 1'7 kb de al menos uno de los eventos bioquímicos detectados en el proyecto	99

* 90% de motivos de factores de transcripción conocidos y cientos de otros nuevos.

** en un tipo celular como mínimo.

- E (*predicted enhancer*): regiones de cromatina abierta ricas en *enhancers*. Presentan una actividad variable en función de la línea celular.
- R (*predicted repressed*): regiones silenciadas.
- TSS: cromatina abierta rica en factores de transcripción que actúan cerca de los promotores y las polimerasas Pol II y Pol III. Tienen dos tipos de actividades y es rico en H3K4me3.
- PF (*predicted promoter flanking region*): se encuentran alrededor de los TSS.
- T (*predicted transcribed region*): cuando cuenta con la modificación H3k36me3, presentan señales de elongación transcripcional.
- WE (*weak enhancer or chromatin cis-regulatory element*).

Se pudo concluir que las características del RNA, de los patrones de metilación y de los factores de transcripción eran distintas según el estado en el que se encontrase la cromatina. Asimismo, los *enhancers* con mayor grado de metilación eran los menos activos y presentaban unos RNA sin colas de poli A.

A continuación, se estudió cómo influían las variaciones de las secuencias *ENCODE*. Para ello, se realizó una comparación de diversos genomas. Las variaciones raras dentro de la secuencia *ENCODE* demostraron afectar a genes codificantes, influyendo en los sitios de unión de los factores de transcripción. En los casos de heterocigosis, los patrones de modificación de histonas eran distintos según el origen materno o paterno de esas histonas. Además, se encontró una correlación positiva entre la actividad del alelo y los ensayos de *ENCODE*, dando una actividad distinta en función del origen. Esto podría deberse a tres factores: a la existencia de una mayor afinidad a alelos accesibles, a la modificación de una histona concreta y a los factores de transcripción.

1.4. Conclusiones

1.4.1 Organización tridimensional del núcleo

En el proyecto *ENCODE* se ha desvelado la increíble complejidad del genoma humano. No solo se han catalogado numerosos elementos reguladores, nuevos transcritos y tipos de RNAs, sino que también ha revelado una intrincada interconexión entre distintas partes del genoma. Este estudio de la interacción entre elementos genómicos tan alejados, otorga un nuevo nivel de complejidad a la regulación de la expresión génica inexplorado hasta ahora. ¿Cómo se regula esta interconexión a nivel tridimensional?, ¿mediante qué mecanismos regiones del genoma rastrean el nucleoplasma para interaccionar con otras?, ¿qué fuerzas gobiernan estos movimientos del DNA?, son algunas de las preguntas que aún esperan respuesta. Estudios realizados en los últimos años han comenzado a revelar la importancia de la envuelta nuclear en la organización tridimensional del genoma (Misteli and Soutoglou, 2009). Así mismo, estudios recientes han puesto de manifiesto que fuerzas generadas por microtúbulos interfásicos sobre la envuelta nuclear producen surcos o fluctuaciones reversibles en la envuelta nuclear, que pueden funcionar como reguladores de la expresión génica al acercar o distanciar distintas partes de la cromatina (Hampoelz et al., 2012; Herrmann, 2013).

1.4.2 Los fármacos epigenómicos. Nuevas estrategias contra el cáncer

Una de las conclusiones que se pueden extraer del proyecto *ENCODE* es que la funcionalidad de los genes codificantes de cada genoma también está asociada a las variaciones que presente fuera de los exones y que la regulación génica tiene que ser entendida como un proceso dinámico. Esto supone un nuevo enfoque en el tratamientos de enfermedades asociadas a la expresión génica y la alteración de la epigenética celular (Dunham et al., 2012).

La desregulación epigenética es una de las causas del cáncer y de su progresión (Ellis et al., 2009). Basados que las modificaciones epigenéticas son reversibles, en los últimos años se han desarrollado fármacos que alteran la epigenética de las células tumorales con objeto de prevenir su malignidad (Ho et al., 2013). Estos fármacos controlan cambios epigenéticos como la acetilación y metilación de histonas o la metilación del DNA. Por ejemplo, en este últimos caso, la hidralacina un inhibidor de la metilación del DNA reactiva la expresión de genes supresores de tumores (Ellis et al., 2009). Estos nuevos tratamientos con fármacos epigenéticos aplicados junto a las terapias ya existentes podrían ser una prometedora forma de combatir el cáncer y abre nuevas expectativas para su investigación, desarrollo y aplicación.

Siempre habrá un antes y un después del proyecto *ENCODE* y aunque, de momento, *ENCODE* solo ha cambiado nuestra forma de ver nuestro propio genoma, es de esperar que en un futuro tenga una importante repercusión sobre la salud humana.

2. AUTORES

(*) Beyond Darwin es el pseudónimo usado por la clase de Genética Molecular de 2º curso de Grado de Biotecnología de la Universidad Pablo de Olavide de Sevilla, para este trabajo.

Los componentes del grupo y autores de esta revisión sobre la reciente publicación en la revista Nature de lo datos del proyecto *ENCODE* sobre el genoma humano, son: Antonio Barral Gil, David Cabrerizo Granados, Irene Perea Romero Lourdes Patricia Román Cano, Mª Carmen Romero Medina, Alejandro Salguero Jiménez, Laura Castro Morales, María Remedios Domínguez Flores, Juan Elías González Correa, Purificación Jiménez Martín, Javier Macías León, Carmen Mangas Corrales, Sergio Sánchez Rivas, José Manuel Marín Morales, Amador Gallardo de los Reyes, Miguel García Ortegón, Javier Méndez Gómez, María Teresa Hato Castro, Alba Jiménez Díaz, Sara Pérez Muñoz, María del Carmen Porcel Sánchez, Pablo Rodríguez Núñez, Lucía Zhu, Clara Rodríguez Fernández, Cristina Ulecia, Jorge Martinez, Lucía Moreno, Cristina Ojeda, Antonio Rafael Ruíz, Isabel Guerrero, Alfonso Alba Bernal, Manuel Arenas Vallejo, Rafael Blanco Domínguez, Belén González Otero, Lucía Morales Cacho, Cristina Gil González, Amparo Martínez Pérez, Sara Molero Rivas, Nuria Morales Puerto, José Luis Sánchez-Trincado López.

Figura. 1. Imagen de la Clase de Genética Molecular 2º Grado Biotecnología (2012-2013) durante la resolución de un examen en grupo.

El trabajo ha sido supervisado por el profesor Rafael R. Daga, responsable de la asignatura Genética Molecular

3. AGRADECIMIENTOS

Nos gustaría agradecer a Paola Gallardo, Alumna Interna del Área de Genética, su contribución al ensamblaje de las distintas partes que constituían este trabajo, así como por sus comentarios y sugerencias.

4. BIBLIOGRAFIA

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. et al. (2012). Landscape of transcription in human cells. *Nature* **489**, 101-8.

Dunham, I. Kundaje, A. Aldred, S. F. Collins, P. J. Davis, C. A. Doyle, F. Epstein, C. B. Frietze, S. Harrow, J. Kaul, R. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y. and Segal, E. (2012). Genomics: ENCODE explained. *Nature* **489**, 52-5.

Ellis, L., Atadja, P. W. and Johnstone, R. W. (2009). Epigenetics in cancer: targeting chromatin modifications. *Mol Cancer Ther* **8**, 1409-20.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R. et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100.

Goldberg, A. D., Allis, C. D. and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell* **128**, 635-8.

Hampoelz, B., Azou-Gros, Y., Fabre, R., Markova, O., Puech, P. H. and Lecuit, T. (2012). Microtubule-induced nuclear envelope fluctuations control chromatin dynamics in Drosophila embryos. *Development* **138**, 3377-86.

Herrmann, H. (2013). Nuclear architecture and dynamics: chromatin, epigenetics, genomics: Review of Genome Organization and Function in the Cell Nucleus: Edited by Karsten Rippe. *Nucleus* **4**, 85-8.

Ho, A. S., Turcan, S. and Chan, T. A. (2013). Epigenetic therapy: use of agents targeting deacetylation and methylation in cancer management. *Onco Targets Ther* **6**, 223-32.

Janicki, S. M., Tsukamoto, T., Salghetti, S. E., Tansey, W. P., Sachidanandam, R., Prasanth, K. V., Ried, T., Shav-Tal, Y., Bertrand, E., Singer, R. H. et al. (2004). From silencing to gene expression: real-time analysis in single cells. *Cell* **116**, 683-98.

Misteli, T. and Soutoglou, E. (2009). The emerging role of nuclear architecture in DNA repair and genome maintenance. *Nat Rev Mol Cell Biol* **10**, 243-54.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K. et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90.

Sanyal, A., Lajoie, B. R., Jain, G. and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* **489**, 109-13.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B. et al. (2012). The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82.

5. GLOSARIO DE TÉRMINOS Y METODOLOGÍA

Promotor: Una región de DNA localizada normalmente aguas arribas de un gen y que determina la transcripción del mismo.

***Enhancer*.** Son secuencias localizadas aguas arribas o abajo de una región promotora y que regulan la transcripción de genes adyacentes, determinando no solo el nivel de expresión de dichos genes, sino, el tejido y el momento del desarrollo en el que éstos se expresarán.

***Splicing* alternativo.** La mayoría de los genes eucariotas están compuestos por exones (regiones codificantes) e intrones (regiones no codificantes). Durante la maduración de un pre-mRNA se eliminan los intrones (*splicing* o adjuste) para producir un transcrito maduro. La generación de diferentes transcritos a partir de un mismo pre-mRNA mediante el adjuste combinatorio de distintos exones para producir distintos mRNA maduros se conoce como *splicing* alternativo.

Sitios hipersensibles a DNasaI (DHSs). La ocupación de tipo no nucleosomal del ADN por parte de proteínas en general y factores

de transcripción en particular puede determinarse mediante ensayos de hipersensibilidad o "footprinting", utilizando para ello la enzima DNasaI (que corta preferentemente dejando extremos pirimidínicos con fosfato en 5' pero es inespecífica de secuencia). Esta estrategia aprovecha que la nucleasa no puede cortar aquellas secuencias que estén físicamente unidas a proteínas, debido al impedimento estérico que suponen éstas al acceso y funcionamiento de la enzima en dichas condiciones. Así es como se generan los llamados "footprints", es decir, los fragmentos que quedan intactos porque las nucleasas son incapaces de degradarlos.

3C-5C . Chromosome Conformation Capture, "Captura de la Conformación de los Cromosomas". El método se basa en el uso de formaldehido para fijar los bucles resultantes de las interacciones entre un fragmento del genoma y otro más distante, formando una estructura conocida como loop. Una vez fijados los dos fragmentos, se degrada el loop, quedando ambos fragmentos ligados intramolecularmente. Las moléculas resultantes de este proceso son amplificadas por PCR, empleando primers específicos. Posteriormente, serán analizados en gel de agarosa. Este método resulta ser de gran utilidad en estudios a pequeña escala, puesto que en el caso contrario se observa que la cantidad de producto de PCR sintetizado es muy pequeño. Por esta razón se ha llevado a cabo una modificación de la técnica, lo que permite obtener más producto analizable: en primer lugar, se realiza una ligación múltiple para copiar los resultados del experimento, y en segundo lugar se amplifican y cuantifican, utilizando *microarrays* o secuenciación cuantitativa. A este nuevo método se le denominó 5C y permite analizar grandes cantidades del genoma, e incluso determinar la posición de los elementos cis o trans.

Inmunoprecipitación de Cromatina Esta metodología se emplea para estudiar la unión de factores de transcripción al ADN y consiste en fijar químicamente los complejos ADN-proteína y fragmentar posteriormente la cadena de DNA, normalmente por sonicación. Posteriormente, se inmunoprecipita las cromatina usando anticuerpos específicos dirigidos contra determinados proteínas de unión al ADN y se analiza el DNA purificado o mediante PCR o secuenciación (*ChIP-seq*).

Método *CAGE* (cap-analysis of gene expression). Permite analizar las zonas de inicio de la transcripción (TSS o transcription start site), así como sus promotores asociados a través del reconocimiento de señales en el extremo 5'.

Locus Control Region (LCR) Región de control de un locus. Son regiones del genoma capaces de controlar el nivel de expresión de ciertos genes.

TSS: Transcription Start Site, Sitio de inicio de la transcripción.

PAS (poliadenylation sites). Sitios de poliadenilación