Perspectiva

¿Qué es un gen, post-ENCODE? Historia y definición actualizada

Marcos B. Gerstein,[1,2,3,9] puede bruce,[2,4] Joel S. Rozowsky,[2] Deyou Zheng,[2] jiang du,[3] Jan O. Korbel,[2,5] Olof Emanuelsson,[6] Zhengdong D Zhang,[2] sherman weissman,[7] y Michael Snyder[2,8]

[1]Programa en Biología Computacional y Bioinformática, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.; [2]Departamento de Biofísica y Bioquímica Molecular, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.; [3]Departamento de Ciencias de la Computación, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.; [4]Centro de Informática Médica, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.; [5]Laboratorio Europeo de Biología Molecular, 69117 Heidelberg, Alemania; [6]Centro de Bioinformática de Estocolmo, Centro Universitario Albanova, Universidad de Estocolmo, SE-10691 Estocolmo, Suecia; [7]Departamento de Genética, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.; [8]Departamento de Biología Molecular, Celular y del Desarrollo, Universidad de Yale, New Haven, Connecticut 06511, EE. UU.

Mientras que la secuenciación del genoma humano nos sorprendió con el número de genes que codifican proteínas que hay, que no cambió fundamentalmente nuestra perspectiva sobre lo que es un gen. Por el contrario, los complejos patrones de regulación dispersa y la transcripción omnipresente descubierto por el proyecto ENCODE, junto con la conservación no génica y la abundancia de ARN no codificantes de genes, han cuestionado la noción del gen. Para ilustrar esto, se revisa la evolución de las definiciones operacionales de un gen durante el siglo pasado, desde los elementos abstractos de la herencia de Mendel y Morgan a los ORF actuales enumeradas en los bancos de datos de secuencias. a continuación, se resumen las conclusiones de ENCODE actuales y proporcionar una metáfora computacional para la complejidad. Finalmente, se propone una actualización de tentativa a la definición de un gen: Un gen es una unión de secuencias genómicas que codifican un conjunto coherente de productos funcionales potencialmente superpuestos. Nuestra definición elude las complejidades de la regulación y la transcripción al eliminar la primera por completo de la definición y argumentar que los productos genéticos funcionales finales (en lugar de transcripciones intermedias) deben usarse para agrupar entidades asociadas con un solo gen. También manifiesta cuán integral es el concepto de función biológica en la definición de genes.

Introducción

La visión clásica de un gen como un elemento discreto en el genoma ha sido sacudida por ENCODE

El consorcio ENCODE completó recientemente su caracterización del 1% del genoma humano mediante varias técnicas experimentales y computacionales de alto rendimiento diseñadas para caracterizar elementos funcionales (The ENCODE Project Consortium 2007). Este proyecto representa un hito importante en la caracterización del genoma humano, y los hallazgos actuales muestran una imagen sorprendente de actividad molecular compleja. Si bien la histórica secuenciación del genoma humano sorprendió a muchos con el pequeño número (en relación con organismos más simples) de genes que codifican proteínas que los anotadores de secuencias podían identificar (~21.000, según la última estimación [ver www.ensembl.org]), ENCODE destacó el número y la complejidad de las transcripciones de ARN que produce el genoma. En este sentido, ENCODE ha cambiado nuestra visión de "qué es un gen" bastante más que la secuenciación de los *Haemophilus influenza* y los genomas humanos lo hicieron (Fleischmann et al. 1995; Lander et al. 2001; Venter et al. 2001). La discrepancia entre nuestra visión anterior del gen centrada en las proteínas y la que se revela por la extensa actividad transcripcional del genoma nos impulsa a reconsiderar ahora qué es un gen. Aquí, revisamos cómo el concepto de gen ha cambiado a lo largo del siglo pasado, resumir el pensamiento actual basado en los últimos hallazgos de ENCODE y proponer una nueva definición de gen actualizada que tenga en cuenta estos hallazgos.

[9]Autor correspondiente.
Correo electrónico Mark.Gerstein@yale.edu ; fax (360) 838-7861.
El artículo está en línea en http://www.genome.org/cgi/doi/10.1101/gr.6339607. Disponible gratuitamente en línea a través de la *Investigación del genoma* Opción de acceso abierto.

Historia del gen, de 1860 hasta justo antes de ENCODE

Definición Década de 1860 a 1900: el gen como una unidad discreta de herencia

El concepto de "gen" ha evolucionado y se ha vuelto más complejo desde que se propuso por primera vez (ver la línea de tiempo en la Fig. 1, póster adjunto). Hay varias definiciones del término, aunque las descripciones iniciales comunes incluyen la capacidad de determinar una característica particular de un organismo y la heredabilidad de esta característica. En particular, la palabra *gene* fue utilizado por primera vez por Wilhelm Johannsen en 1909, basado en el concepto desarrollado por Gregor Mendel en 1866 (Mendel 1866). La palabra era un derivado de *pangen*, que fue utilizado por Hugo De Vries para las entidades involucradas en la pangénesis, el mecanismo hipotético de la herencia de Darwin (Heimans 1962). Johnannsen llama un gen los "especiales condiciones, fundaciones y determinantes que están presentes [en el gametos] en formas únicas, separados y por lo tanto independientes [por que] muchas características del organismo se especifican" (Johannsen 1909, p. 124). La etimología del término deriva del griego *génesis* ("nacimiento") o *genos* ("origen"). La palabra relacionada *genética* fue utilizado por el genetista William Bateson en 1905 (http://www.jic.ac.uk/corporate/about/bateson.htm).

Mendel demostró que cuando se crían plantas, algunos rasgos como la altura o el color de las flores no aparecen mezclados en sus plantas.

Figura 1.(Afiche adjunto) Cronología de la historia del término "gen". Un término inventado hace casi un siglo, "gen", con su seductora ortografía simple, se ha convertido en un concepto central en biología. Dado un significado específico en su acuñación, esta palabra se ha convertido en algo complejo y esquivo a lo largo de los años, lo que refleja nuestro conocimiento cada vez mayor en genética y en ciencias de la vida en general. Los sorprendentes descubrimientos realizados en el Proyecto ENCODE, como muchos antes que enriquecieron significativamente el significado de este término, son precursores de otra ola de cambio en nuestra comprensión de lo que es un gen.

primavera, es decir, estos rasgos se transmiten como entidades distintas y discretas (Mendel 1866). Su trabajo también demostró que las variaciones en los rasgos eran causadas por variaciones en los factores hereditarios (o, en la terminología actual, el fenotipo es causado por el genotipo). Fue solo después de que Carl Correns, Erich von Tschermak-Seysenegg y Hugo De Vries repitieron y redescubrieron el trabajo de Mendel en 1900 que realmente comenzó un trabajo adicional sobre la naturaleza de la unidad de herencia (Tschermak 1900; Vries 1900; Rheinberger 1995).

Definición Década de 1910: el gen como un locus distinto

En el próximo desarrollo importante, el genetista estadounidense Thomas Hunt Morgan y sus estudiantes estaban estudiando la segregación de mutaciones en*Drosophila melanogaster*. Pudieron explicar sus datos con un modelo en el que los genes están dispuestos linealmente y su capacidad de cruce es proporcional a la distancia que los separaba. El primer mapa genético se creó en 1913 (Sturtevant 1913), y Morgan y sus estudiantes publicaron*El mecanismo de la herencia mendeliana*en 1915 (Morgan et al. 1915). Para los primeros genetistas, un gen era una entidad abstracta cuya existencia se reflejaba en la forma en que se transmitían los fenotipos entre generaciones. La metodología utilizada por los primeros genetistas involucraba mutaciones y recombinación, por lo que el gen era esencialmente un locus cuyo tamaño estaba determinado por mutaciones que inactivaban (o activaban) un rasgo de interés y por el tamaño de las regiones recombinantes. El hecho de que el enlace genético correspondiera a ubicaciones físicas en los cromosomas fue demostrado más tarde, en 1929, por Barbara McClintock, en sus estudios citogenéticos sobre el maíz (McClintock 1929).

Definición Década de 1940: el gen como modelo para una proteína

Beadle y Tatum (1941), quienes estudiaron*Neurospora*metabolismo, descubrió que las mutaciones en los genes podrían causar defectos en los pasos de las vías metabólicas. Esto se expresó como el punto de vista de "un gen, una enzima", que luego se convirtió en "un gen, un polipéptido". Desde este punto de vista, el gen se considera implícitamente como la información detrás de las moléculas individuales en una vía bioquímica. Esta visión se volvió progresivamente más explícita y mecanicista en décadas posteriores.

Definición Década de 1950: el gen como molécula física

El hecho de que la herencia tiene una base física y molecular fue demostrado por la observación de que los rayos X pueden causar mutaciones (Muller 1927). La demostración de Griffith (1928) de que algo en virulento pero muerto*Neumococo*cepas podrían ser absorbidas por organismos vivos no virulentos*Neumococo*y transformarlos en bacterias virulentas fue una prueba más en este sentido. Posteriormente se demostró que esta sustancia podía ser destruida por la enzima DNasa (Avery et al. 1944). En 1955, Hershey y Chase establecieron que la sustancia realmente transmitida por el bacteriófago a su progenie es el ADN y no la proteína (Hershey y Chase 1955). Además, la idea de que el producto de un gen es una sustancia difusible subyace a la prueba de complementación que se utilizó para definir los genes.

en los primeros años de la bacteriología. Una visión práctica del gen era la del cistrón, una región del ADN definida por mutaciones que en *trans*no podían complementarse genéticamente (Benzer 1955).

Definición 1960: gen como código transcrito

Fue la solución de la estructura tridimensional del ADN por Watson y Crick en 1953 (Watson y Crick 1953) lo que explicó cómo el ADN podría funcionar como la molécula de la herencia. El apareamiento de bases explicaba cómo se podía copiar la información genética, y la existencia de dos hebras explicaba cómo los errores ocasionales en la replicación podían conducir a una mutación en una de las copias hijas de la molécula de ADN.

A partir de la década de 1960, la biología molecular se desarrolló a un ritmo acelerado. La transcripción de ARN de las secuencias codificantes de proteínas se tradujo utilizando el código genético (resuelto en 1965 por Nirenberg et al. [1965] y Söll et al. [1965]) en una secuencia de aminoácidos. Francis Crick (1958) resumió el flujo de información en la expresión génica desde el ácido nucleico hasta la proteína (los comienzos del "Dogma Central"). Sin embargo, hubo algunas excepciones inmediatas a esto: se sabía que algunos genes no codifican proteínas sino moléculas de ARN funcionales como el ARNr y el ARNt. Además, en los virus de ARN, el gen está hecho de ARN. La visión molecular del gen que se desarrolló durante la década de 1960 se puede resumir en términos generales como un código que reside en el ácido nucleico que da lugar a un producto funcional.

Definición Décadas de 1970 a 1980: gen como patrón de secuencia de marco de lectura abierto (ORF)

El desarrollo de clonación y técnicas en la década de 1970, en combinación con el conocimiento del código genético secuenciación, revolucionó el campo de la biología molecular, proporcionando una gran cantidad de información sobre cómo los genes se organizan y se expresaron. El primer gen que se secuenció fue del bacteriófago MS2, que también fue el primer organismo que se secuenció por completo (Fiers et al. 1971, 1976). El desarrollo paralelo de herramientas computacionales condujo a algoritmos para la identificación de genes basados en sus características de secuencia (p. ej., para una revisión, véase Rogic et al. 2001). En muchos casos, se podría usar una secuencia de ADN para inferir la estructura y función del gen y sus productos. Esta situación creó un nuevo concepto de "gen nominal", que se define por su secuencia predicha en lugar de un locus genético responsable de un fenotipo (Griffiths y Stotz 2006). La identificación de la mayoría de los genes en genomas secuenciados se basa en su similitud con otros genes conocidos o en la firma estadísticamente significativa de una secuencia codificante de proteínas. En muchos casos, el gen se identificó efectivamente como un ORF anotado en el genoma (Doolittle 1986).

Definición 1990s-2000s: entidad genómica anotada, enumerada en los bancos de datos (vista actual, pre-ENCODE)

La definición actual de un gen utilizada por organizaciones científicas que anotan genomas aún se basa en la vista de secuencia. Por lo tanto, un gen fue definido por la Organización de Nomenclatura del Genoma Humano como "un segmento de ADN que contribuye al fenotipo/función. En ausencia de una función demostrada, un gen puede caracterizarse por secuencia, transcripción u homología" (Wain et al. 2002). Recientemente, el Sequence Ontology Consortium supuestamente llamó al gen una "región localizable de secuencia genómica, correspondiente a una unidad de herencia, que está asociada con

regiones reguladoras, regiones transcritas y/u otras regiones de secuencias funcionales" (Pearson 2006).

La secuenciación de primero el*Haemophilus influenza*y luego el genoma humano (Fleischmann et al. 1995; Lander et al. 2001; Venter et al. 2001) llevaron a una explosión en la cantidad de secuencias a las que se podían aplicar definiciones como las anteriores. De hecho, hubo un gran interés popular en contar el número de genes en varios organismos. Este interés cristalizó originalmente con la apuesta de Gene Sweepstake sobre el número de genes en el genoma humano, que recibió una amplia cobertura mediática (Wade 2003).

Se ha señalado que estas enumeraciones exageran los genes codificadores de proteínas tradicionales. En particular, cuando se informó el número de genes presentes en el genoma humano en 2003, se reconoció que se sabía muy poco sobre los genes que codifican el ARN, de modo que el número dado era el de los genes que codifican proteínas. La visión de Ensembl del gen se resumió específicamente en las reglas del Gene Sweepstake de la siguiente manera: "todos los transcritos empalmados alternativamente pertenecen al mismo gen, incluso si las proteínas que se producen son diferentes". (http://web.archive.org/web/20050627080719/www.ensembl.org/Genesweep/).

Una corriente metáfora computacional: Genes como "subrutinas" en el sistema operativo genómico

Dado que contar genes en el genoma es un esfuerzo computacional a gran escala y que los genes se ocupan fundamentalmente del procesamiento de información, el léxico de la informática, naturalmente, se ha aplicado cada vez más para describirlos. En particular, las personas en la comunidad de biología computacional han utilizado la descripción de un lenguaje formal para describir la estructura de los genes de la misma manera que se utilizan las gramáticas para describir los programas informáticos, con una sintaxis precisa de regulación, exones e intrones aguas arriba. (Searls 1997, 2001, 2002). Además, una metáfora que es cada vez más popular para describir los genes es pensar en ellos en términos de subrutinas en un gran sistema operativo (SO). Es decir, en la medida en que los nucleótidos del genoma se agrupan en un código que se ejecuta a través del proceso de transcripción y traducción, el genoma puede pensarse como un sistema operativo para un ser vivo. Los genes son entonces subrutinas individuales en este sistema general que se llaman repetidamente en el proceso de transcripción.

Cuestiones problemáticas con la definición actual de un gen

Hay una serie de aspectos problemáticos de la definición actual de un gen, tal como se aplica al genoma humano, que se discuten a continuación. Varias complicaciones adicionales se resumen en la Tabla 1.

1. Regulación de genes

Jacob y Monod (1961), en su estudio de la*laca*operón de *Escherichia coli*, Proporcionado un paradigma para el mecanismo de regulación del gen: Consistía en una región de ADN que consiste en secuencias que codifican una o más proteínas, una secuencia "promotor" para la unión de la ARN polimerasa, y una secuencia de "operador" a la que se unen los genes reguladores. Más tarde, se descubrió que existían otras secuencias que podrían afectar prácticamente todos los aspectos de la regulación génica, desde la transcripción hasta la degradación del ARNm y modificación post-traduccional. Tales secuencias podrían residir dentro de la secuencia de codificación así como en las regiones flanqueantes y, en el caso de los potenciadores y elementos relacionados, muy lejos de la secuencia de codificación. Aunque funcionalmente necesarios para la expresión del producto génico, los elementos reguladores, especialmente los distantes, hicieron problemático el concepto del gen como un locus genético compacto.

La regulación es parte integral de muchas definiciones actuales del gen. En particular, una definición de libro de texto actual de un gen en términos moleculares es la secuencia de ácido nucleico completa que es necesaria para la síntesis de un polipéptido funcional (o ARN) (Lodish et al. 2000). Si eso implica una síntesis debidamente regulada, las secuencias de ADN en un gen incluirían no solo aquellas que codifican el pre-ARNm y sus regiones de control flanqueantes, sino también potenciadores. Además, muchos potenciadores están distantes a lo largo de la secuencia de ADN, aunque en realidad están bastante cerca debido a la estructura tridimensional de la cromatina.

2. Genes superpuestos y empalmados

superposición

A medida que se secuenciaron los genes, los ARNm y, finalmente, los genomas completos, el modelo de operón simple resultó ser aplicable solo a los genes de procariotas y sus fagos. Los eucariotas eran diferentes en muchos aspectos, incluida la organización genética y el flujo de información. El modelo de genes como unidades hereditarias que no se superponen y son continuos demostró ser incorrecto mediante el mapeo preciso de las secuencias de codificación de los genes. De hecho, se ha descubierto que algunos genes se superponen entre sí, compartiendo la misma secuencia de ADN en un marco de lectura diferente o en la hebra opuesta. La estructura discontinua de los genes permite potencialmente que un gen esté completamente contenido dentro del intrón de otro, o que un gen se superponga con otro en la misma hebra sin compartir ningún exón o elemento regulador.

empalme

El empalme se descubrió en 1977 (Berget et al. 1977; Chow et al. 1977; Gelinas y Roberts 1977). Pronto quedó claro que el gen no era una simple unidad de herencia o función, sino más bien una serie de exones, que codificaban, en algunos casos, dominios discretos de proteínas y estaban separados por largos tramos no codificantes llamados intrones. Con el empalme alternativo, un locus genético podría codificar múltiples transcripciones de ARNm diferentes. Este descubrimiento complicó radicalmente el concepto de gen. Por ejemplo, en la secuenciación del genoma, Celera definió un gen como "un locus de exones co-transcritos" (Venter et al. 2001), y la página web Gene Sweepstake de Ensembl originalmente definió un gen como "un conjunto de transcritos conectados, " donde "conectado" significa compartir un exón (http://web.archive.org/web/20050428090317/www.ensembl.org/Genesweep).

Trans-*empalme*

el fenómeno de*trans*-el empalme (ligadura de dos moléculas de ARNm separadas) complicó aún más nuestra comprensión (Blumenthal 2005). Hay ejemplos de transcripciones del mismo gen, o de la hebra de ADN opuesta, o incluso de otro cromosoma, que se unen antes de empalmarse. Claramente, el concepto clásico del gen como "un locus" ya no se aplica a estos productos génicos cuyas secuencias de ADN están muy separadas a lo largo del genoma.

Tabla 1.Fenómenos que complican el concepto de gen

Fenómeno	Descripción	Tema
Ubicación y estructura de genes genes intronic	Un gen existe dentro de un intrón de otro (Henikoff et al. 1986)	Dos genes en el mismo locus
Genes con marcos de lectura superpuestos	Una región de ADN puede codificar para dos diferentes productos proteicos en diferentes marcos de lectura (Contreras et al. 1977)	No hay correspondencia uno a uno entre el ADN y secuencia de proteínas
Potenciadores, silenciadores	Elementos reguladores distantes (Spilianakis et al. 2005)	Las secuencias de ADN que determinan la expresión pueden ser muy separados unos de otros en el genoma. Relación de muchos a muchos entre los genes y sus potenciadores.
variación estructural		
Elementos móviles	El elemento genético aparece en nuevas ubicaciones sobre generaciones (McClintock 1948)	Un elemento genético puede no ser constante en su localización
Reordenamientos genéticos/variantes estructurales	Reordenamiento o empalme del ADN en células somáticas da como resultado muchos productos genéticos alternativos (Early et al. 1980)	La estructura del gen no es hereditaria, o la estructura puede diferir entre individuos o células/tejidos
Variantes de número de copia	El número de copias de genes/elementos reguladores puede difieren entre individuos (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005)	Los elementos genéticos pueden diferir en su número.
Epigenética y estructura cromosómica		
Modificaciones epigenéticas, impronta	La información heredada puede no ser una secuencia de ADN (p. ej., Dobrovic et al. 1988); la expresión de un gen depende de si es de origen paterno o materno (Sager y Kitchin 1975)	El fenotipo no está determinado estrictamente por genotipo
Efecto de la estructura de la cromatina	Estructura de la cromatina, que sí influye en los genes. expresión, solo vagamente asociada con secuencias de ADN particulares (Paul 1972)	La expresión génica depende del empaquetamiento del ADN. La secuencia de ADN no es suficiente para predecir el producto génico.
Eventos postranscripcionales		
Empalme alternativo de ARN	Una transcripción puede generar múltiples ARNm, dando como resultado diferentes productos proteicos (Berget et al. 1977; Gelinas y Roberts 1977) Marcos de	Múltiples productos de un locus genético; información en el ADN no relacionada linealmente con la de la proteína
Productos empalmados alternativamente con alternativas marcos de lectura	lectura alternativos del tumor INK4a El gen supresor codifica dos proteínas no relacionadas (Quelle et al. 1995)	Dos productos de empalme alternativos de un pre-ARNm producir productos proteicos sin una secuencia en común
ARN*trans*-empalme, homotípico*trans*-empalme	secuencias de ADN distantes pueden codificar para las transcripciones se ligó en diversas combinaciones (Borst 1986). Dos transcritos idénticos de un gen pueden *trans* -empalme para generar un ARNm donde se repite la misma secuencia de exón (Takahara et al. 2000).	Una proteína puede resultar de la combinación información codificada en múltiples transcripciones
edición de ARN	El ARN se modifica enzimáticamente (Eisen 1988)	La información del ADN no está codificada. directamente en la secuencia de ARN
Eventos postraduccionales Empalme de proteínas, poliproteínas virales	El producto proteico se escinde por sí mismo y puede generar múltiples productos funcionales (Villa-Komaroff et al. 1975)	Sitios de inicio y finalización de la proteína no determinados por codigo genetico
Proteína*trans*-empalme	Proteínas distintas pueden empalmarse juntas en el ausencia de un*trans*-transcripción empalmada (Handa et al. 1996)	Sitios de inicio y finalización de la proteína no determinados por codigo genetico
modificación de proteínas	La proteína se modifica para alterar la estructura y función del producto final (Wold 1981)	La información del ADN no está codificada. directamente en la secuencia de proteínas
Pseudogenes y retrogenes		
Retrogenes	Un retrogén se forma a partir de la transcripción inversa. de ARNm de su gen padre (Vanin et al. 1980) y mediante la inserción del producto de ADN en un genoma	Flujo de información de ARN a ADN
Pseudogenes transcritos	Se transcribe un pseudogen (Zheng et al. 2005, 2007)	Actividad bioquímica de supuestamente muertos elementos

Finalmente, varios estudios recientes han destacado un fenómeno denominado quimerismo en tándem, en el que dos genes consecutivos se transcriben en un solo ARN (Akiva et al. 2006; Parra et al. 2006). La traducción (después de corte y empalme) de tales ARN puede conducir a una nueva, proteína fusionada, que tiene partes de ambas proteínas originales.

3. Genes parásitos y móviles

Un desafío a nuestro concepto del gen ha sido el de la contienda o gen parasitaria. La idea propuesta por primera vez por Richard Dawkins es que la unidad de la evolución no es el organismo pero el gen (Dawkins, 1976). Los organismos son herramientas sólo que los genes utilizan para repre-

licarse ellos mismos. El concepto de Dawkins del optimon (o selecton) es una unidad de ADN que sobrevive a la recombinación durante suficientes generaciones para ser seleccionadas juntas.

El término parasitaria ciertamente parece apropiado para los transposones, cuya única función es la de replicarse a sí mismos y que no proporcionan ningún beneficio obvio para el organismo. Los transposones pueden cambiar su ubicación, además de copiarse a sí mismas mediante la escisión, la recombinación o la transcripción inversa. Fueron descubiertos por primera vez en la década de 1930 en el maíz y más tarde se encontró que existen en todas las ramas de la vida, incluyendo los seres humanos (McClintock 1948). Los transposones han cambiado nuestra visión del gen mediante la demostración de que un gen no es fijo en su lugar.

4. La gran cantidad de “ADN basura” bajo selección

El "concepto de patrón de secuencia ORF" del gen tal como existió desde la década de 1980 en adelante dejó en claro que había una gran extensión de elementos no génicos en los genomas eucariotas, particularmente en el genoma humano. En ausencia de conocimiento de una función para estas regiones, algunos propusieron que carecían de una función y usaron la etiqueta de "ADN basura" (Ohno 1972). Esto fue subrayado por la secuenciación posterior del genoma humano, donde se demostró que solo el 1,2% de las bases de ADN codifican exones (Lander et al. 2001; Venter et al. 2001). Sin embargo, algunos de los primeros experimentos piloto de genómica funcional en los cromosomas 21 y 22 indicaron que se transcribieron cantidades apreciables del ADN supuestamente basura (Kapranov et al. 2002; Rinn et al. 2003). Además, la comparación del humano, perro, ratón,~5% bajo selección negativa desde la divergencia de estas especies (Waterston et al. 2002; Lindblad-Toh et al. 2005).

La visión moderna de ENCODE de la actividad del genoma disperso

Como se describió anteriormente, antes del advenimiento del proyecto ENCODE, había una serie de aspectos de los genes que eran muy complicados, pero gran parte de esta complejidad, en cierto sentido, se barrió debajo de la alfombra y realmente no afectó la definición fundamental de un gen. La experiencia del proyecto ENCODE, particularmente el mapeo de la actividad transcripcional y la regulación utilizando matrices de mosaico, ha ampliado estos aspectos desconcertantes y confusos de los genes, llevándolos al frente, donde uno tiene que lidiar más directamente con ellos en relación con la definición de que es un gen

Lo que muestran los experimentos ENCODE: redes de transcritos largos y regulación dispersa

Transcripción sin anotaciones

Un primer hallazgo del consorcio ENCODE que ha reproducido resultados anteriores (Bertone et al. 2004; Cheng et al. 2005) es que una gran cantidad de ADN, no anotado como genes conocidos, se transcribe en ARN (The ENCODE Project Consortium 2007) . Estas nuevas regiones transcritas se denominan normalmente TAR (es decir, regiones transcripcionalmente activas) y transfrags. Si bien la mayor parte del genoma parece transcribirse a nivel de transcripciones primarias, solo aproximadamente la mitad de la transcripción procesada (empalmada) detectada en todas las líneas celulares y condiciones mapeadas se anota actualmente como genes.

TSS no anotados y alternativos

Una segunda observación es que hay un gran número de sitios de inicio de la transcripción unannotated (DST) identificados por cualquiera de secuenciación del extremo 5 de los ARNm transcritos o la asignación de factores de transcripción del promotor asociado a través de chip-chip o chip-PET (El ENCODE Proyecto Consorcio 2007). Además, el consorcio descubrió que muchos genes de proteínas conocidos tienen TSS alternativos que a veces están > 100 kb aguas arriba del sitio de inicio de la transcripción anotado. En particular, Denoeud et al. (2007) realizaron 5 amplificaciones rápidas de extremos de cDNA (RACE) en los 399 loci codificadores de proteínas bien caracterizados contenidos en las regiones ENCODE. El cebador RACE se seleccionó de un exón de 5 que se compartió entre la mayoría de las transcripciones anotadas de cada locus, y los productos RACE se hibridaron en matrices y se mapearon. Descubrieron que más de la mitad de los loci tenían un sitio de inicio de transcripción alternativo aguas arriba del sitio conocido en al menos uno de los 12 tejidos analizados. Algunos de estos TSS distales utilizaron el promotor de un locus génico completamente diferente (es decir, comparten el mismo sitio de inicio de la transcripción). La importancia de este descubrimiento es que el TSS alternativo para algunas de estas transcripciones comenzó dos o tres loci de genes aguas arriba del locus del que se seleccionó el cebador RACE. Por lo tanto, algunas isoformas alternativas son transcripciones que abarcan múltiples loci de genes. (En la Fig. 2 se muestra un esquema de dibujos animados). Muchas de las isoformas alternativas codifican la misma proteína que difiere solo en sus 5 regiones no traducidas (UTR). Algunos de estos TSS distales utilizaron el promotor de un locus génico completamente diferente (es decir, comparten el mismo sitio de inicio de la transcripción). La importancia de este descubrimiento es que el TSS alternativo para algunas de estas transcripciones comenzó dos o tres loci de genes aguas arriba del locus del que se seleccionó el cebador RACE. Por lo tanto, algunas isoformas alternativas son transcripciones que abarcan múltiples loci de genes. (En la Fig. 2 se muestra un esquema de dibujos animados). Muchas de las isoformas alternativas codifican la misma proteína y difieren solo en sus 5 regiones no traducidas (UTR). Algunos de estos TSS distales utilizaron el promotor de un locus génico completamente diferente (es decir, comparten el mismo sitio de inicio de la transcripción). La importancia de este descubrimiento es que el TSS alternativo para algunas de estas transcripciones comenzó dos o tres loci de genes aguas arriba del locus del que se seleccionó el cebador RACE. Por lo tanto, algunas isoformas alternativas son transcripciones que abarcan múltiples loci de genes. (En la Fig. 2 se muestra un esquema de dibujos animados). Muchas de las isoformas alternativas codifican la misma proteína y difieren solo en sus 5 regiones no traducidas (UTR). algunas isoformas alternativas son transcripciones que abarcan múltiples loci de genes. (En la Fig. 2 se muestra un esquema de dibujos animados). Muchas de las isoformas alternativas codifican la misma proteína que difiere solo en sus 5 regiones no traducidas (UTR). algunas isoformas alternativas son transcripciones que abarcan múltiples loci de genes. (En la Fig. 2 se muestra un esquema de dibujos animados). Muchas de las isoformas alternativas codifican la misma proteína y difieren solo en sus 5 regiones no traducidas (UTR).

Más empalmes alternativos

Teniendo en cuenta estos hallazgos, el equipo de La Habana en el Instituto Sanger produjo la anotación GENCODE bien seleccionada (Harrow et al. 2006). No han encontrado que el número de loci de genes codificadores de proteínas conocidos haya aumentado significativamente con el tiempo. Por el contrario, ha aumentado el número de isoformas alternativas anotadas por locus. (La anotación GENCODE actualmente contiene un promedio de 5,4 transcripciones por locus). Por lo tanto, mientras que parte de la gran cantidad de transcripción nueva y sin anotaciones podría corresponder a loci de genes codificantes de proteínas completamente nuevos, es probable que la mayor parte corresponda a segmentos de transcripciones empalmadas alternativamente sin anotaciones que involucran loci de genes conocidos o a ARN no codificantes completamente nuevos.

Regulación dispersa

Como se esquematiza en la Figura 2B, el proyecto ENCODE ha proporcionado evidencia de una regulación dispersa que se extiende por todo el genoma (The ENCODE Project Consortium 2007). Por otra parte, los sitios de regulación para un gen dado no están en fases anteriores necesariamente directamente de ella y pueden, de hecho, estar situados lejos en el cromosoma, más cerca de otro gen. Si bien la unión de muchos factores de transcripción parece cubrir todo el genoma, no se organiza de acuerdo con expectativas aleatorias simples y tiende a agruparse en "bosques" reguladores ricos y "desiertos" pobres (Zhang et al. 2007).

Además, parece que algunos de los elementos reglamentarios pueden en realidad ser transcritos. En un modelo génico convencional y conciso, un elemento de ADN (p. ej., promotor, potenciador y aislante) que regula la expresión génica no se transcribe y, por lo tanto, no forma parte de la transcripción de un gen. Sin embargo, muchos estudios iniciales han descubierto en casos específicos que los elementos reguladores pueden residir en regiones transcritas, como el*lac*operador (Jacob y Monod 1961), un potenciador para regular el gen de la beta-globina (Tuan et al. 1989), y el sitio de unión al ADN del factor YY1 (Shi et al. 1991). El proyecto ENCODE y otras experiencias recientes de ChIP-chip

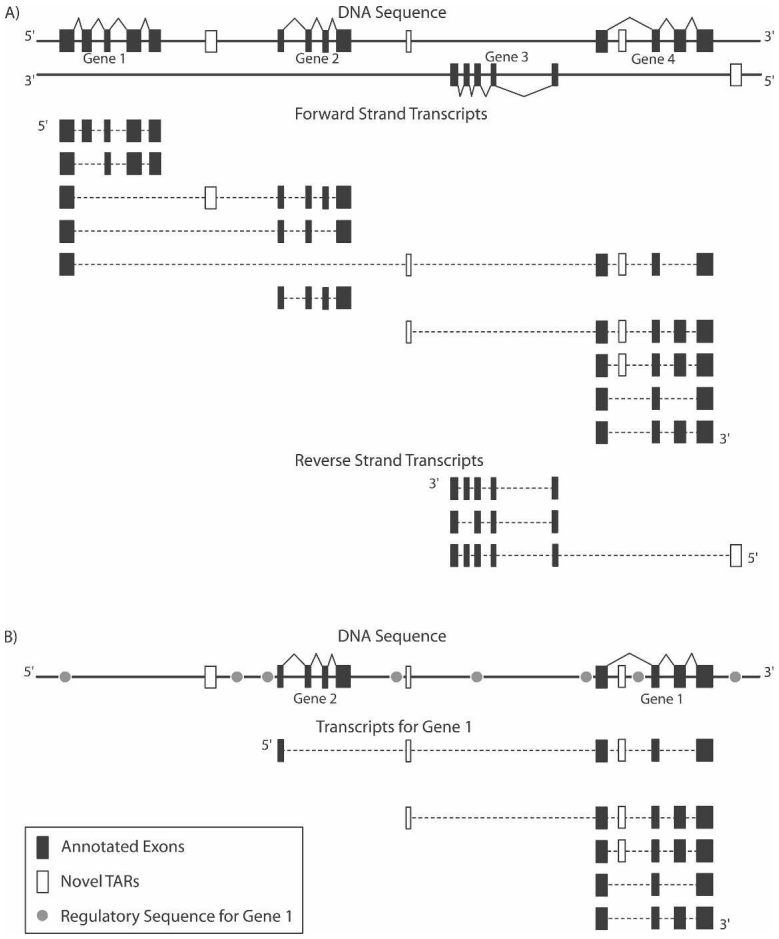

Figura 2. Complejidad biológica revelada por ENCODE. (*UN*) Representación de una región genómica típica región que representa la complejidad de las transcripciones en el genoma. (*Cima*) Secuencia de ADN con exones de genes anotados (rectángulos negros) y TAR nuevos (rectángulos huecos). (*Abajo*) Las diversas transcripciones que surgen de la región de las cadenas directa e inversa. (Líneas punteadas) Intrones empalmados. La anotación de genes convencional representaría solo una parte de las transcripciones provenientes de los cuatro genes en la región (indicada). Los datos del proyecto ENCODE revelan que hay muchas transcripciones que se extienden a lo largo de múltiples loci de genes, algunas de las cuales utilizan sitios de inicio de transcripción 5 distales. (*B*) Representación de las diversas secuencias reguladoras identificadas para un gen diana. Para el Gen 1, mostramos todas las transcripciones de los componentes, incluidas muchas isoformas novedosas, además de todas las secuencias identificadas para regular el Gen 1 (círculos grises). Observamos que algunas de las secuencias potenciadoras son en realidad promotores de nuevas isoformas de empalme. Además, algunas de las secuencias reguladoras del gen 1 podrían estar más cerca de otro gen, y el objetivo se identificaría erróneamente si se eligiera únicamente en función de la proximidad.

Los experimentos han proporcionado evidencia a gran escala de que el modelo de gen conciso puede ser demasiado simple, y muchos elementos reguladores en realidad residen dentro del primer exón, los intrones o el cuerpo completo de un gen (Cawley et al. 2004; Euskirchen et al. 2004; Kim et al. 2005; The ENCODE Project Consortium 2007; Zhang et al. 2007).

Genético versus intergénico: ¿Hay alguna distinción?

En general, los experimentos ENCODE han revelado un rico tapiz de transcripción que involucra empalmes alternativos, cubriendo el genoma en una red compleja de transcripciones. De acuerdo con las definiciones tradicionales, los genes son regiones unitarias de secuencia de ADN, separados unos de otros. ENCODE revela que si uno intenta definir un gen sobre la base de transcripciones superpuestas compartidas, entonces muchos loci de genes distintos anotados se unen en regiones genómicas más grandes. Una implicación obvia de los resultados de ENCODE es que hay menos distinción que hacer entre regiones génicas e intergénicas. Los genes ahora parecen extenderse a lo que una vez se llamó espacio intergénico, con transcripciones recién descubiertas que se originan en sitios reguladores adicionales. Además, hay mucha actividad entre los genes anotados en el espacio intergénico. Dos fuentes bien caracterizadas pueden contribuir a esto, los ARN no codificantes de proteínas (ncRNA) transcritos y los pseudogenes transcritos, y una fracción apreciable de estos elementos transcritos están bajo restricción evolutiva. Varios de estos pseudogenes y genes ncRNA transcritos están, de hecho, ubicados dentro de intrones de genes que codifican proteínas. Uno no puede simplemente ignorar estos componentes dentro de los intrones porque algunos de ellos pueden influir en la expresión de sus genes huésped, ya sea directa o indirectamente.

ARN no codificantes

Las funciones de los genes ncRNA son bastante diversas, incluida la regulación de genes (p. ej., miRNA), el procesamiento de RNA (p. ej., sno-RNA) y la síntesis de proteínas (tRNA y rRNA) (Eddy 2001; Mattick y Makunin 2006). Debido a la falta de codones y, por lo tanto, marcos de lectura abiertos, los genes ncRNA son difíciles de identificar y, por lo tanto, probablemente solo se conoce una fracción de los ncRNA funcionales en humanos hasta la fecha, con la excepción de los que tienen la estructura evolutiva y/o estructural más fuerte. restricciones, que pueden identificarse computacionalmente a través de análisis de plegamiento y coevolución del ARN (p. ej., miARN que muestran estructuras precursoras características en forma de horquilla, o ARNnc en complejos de ribonucleoproteína que, en combinación con péptidos, forman estructuras secundarias específicas) (Washietl et al. 2005, 2007; Pedersen et al. 2006). Sin embargo, el ejemplo de la gran de 17 kb *XIST*El gen involucrado en la compensación de dosis muestra que los ncRNA funcionales pueden expandirse significativamente más allá de las regiones restringidas e identificables computacionalmente (Chureau et al. 2002; Duret et al. 2006).

También es posible que los productos de ARN en sí mismos no tengan una función, sino que reflejen o sean importantes para un proceso celular particular. Por ejemplo, la transcripción de una región reguladora podría ser importante para la accesibilidad de la cromatina para la unión del factor de transcripción o para la replicación del ADN. Tal transcripción se ha encontrado en la región de control del locus (LCR) del locus de la beta-globina, y se ha sugerido que la actividad de la polimerasa es importante para la replicación del ADN en*E. coli*. Alternativamente, la transcripción podría reflejar la actividad no específica de una región en particular, por ejemplo, el reclutamiento de la polimerasa en los sitios reguladores. En cualquiera de estos escenarios, las transcripciones en sí mismas carecerían de una función y es poco probable que se conserven.

pseudogenes

Los pseudogenes son otro grupo más de componentes genómicos "misteriosos" que a menudo se encuentran en los intrones de los genes o en el espacio intergénico (Torrents et al. 2003; Zhang et al. 2003). Se derivan de genes funcionales (mediante retrotransposición o duplicación) pero han perdido las funciones originales de sus genes parentales (Balakirev y Ayala 2003). A veces, oscilando entre vivos y muertos, los pseudogenes pueden influir en la estructura y función del genoma humano. Su prevalencia (tanto como los genes que codifican proteínas) y su estrecha similitud con los genes funcionales ya han confundido la anotación de genes. Recientemente, también se ha encontrado que una fracción significativa (hasta el 20 %) de ellos están transcripcionalmente vivos, lo que sugiere que se debe tener cuidado al usar la expresión como evidencia para localizar genes (Yano et al. 2004; Harrison et al. 2005). , Zheng et al. 2005, 2007; Frith et al. 2006). De hecho, algunos de los nuevos TAR se pueden atribuir a la transcripción de pseudogenes (Bertone et al. 2004; Zheng et al. 2005). En unos pocos casos sorprendentes, se encontró que un pseudogen de ARN o al menos una parte del mismo estaba empalmado con la transcripción de su gen vecino para formar una transcripción quimérica gen-pseudogen. Estos hallazgos agregan una capa adicional de complejidad para establecer la estructura precisa de un locus genético. Además, también se han descubierto transcripciones de pseudogenes funcionales en células eucariotas, como las neuronas del caracol. se encontró un ARN pseudogene o al menos una pieza de la misma para ser empalmado con la transcripción de su gen vecinas para formar una transcripción quimérico gen-pseudogene. Estos resultados se suman una capa adicional de complejidad para establecer la estructura exacta de un locus del gen. Además, las transcripciones pseudogene funcionales también se han descubierto en las células eucariotas, tales como las neuronas del caracol Se descubrió que un pseudogen de ARN o al menos una parte del mismo estaba empalmado con la transcripción de su gen vecino para formar una transcripción quimérica gen-pseudogen. Estos hallazgos agregan una capa adicional de complejidad para establecer la estructura precisa de un locus genético. Además, las transcripciones pseudogene funcionales también se han descubierto en las células eucariotas, tales como las neuronas del caracol *Lymnaea stagnalis*(Korneev et al. 1999). También, curiosamente, el ser humano*XIST*El gen mencionado anteriormente en realidad surge del cuerpo muerto de un pseudogen (Duret et al. 2006). La transcripción de pseudogenes y el límite borroso entre genes y pseudogenes (Zheng y Gerstein 2007) enfatiza una vez más que la naturaleza funcional de muchos TAR nuevos debe resolverse mediante futuros experimentos bioquímicos o genéticos (para una revisión, consulte Gingeras 2007).

Elementos restringidos

Las regiones intergénicas no codificantes contienen una gran fracción de elementos funcionales identificados al examinar los cambios evolutivos en múltiples especies y dentro de la población humana. El proyecto ENCODE observó que solo el 40% de las bases restringidas evolutivamente estaban dentro de los exones codificadores de proteínas o sus regiones no traducidas asociadas (The ENCODE Project Consortium 2007). La resolución de los elementos restringidos identificados por el análisis multiespecies en el proyecto ENCODE es muy alta, identificando secuencias tan pequeñas como 8 bases (con una mediana de 19 bases) (The ENCODE Project Consortium 2007). Esto sugiere que los loci de codificación de proteínas pueden verse como un grupo de pequeños elementos restringidos dispersos en un mar de secuencias sin restricciones. Aproximadamente otro 20% de los elementos restringidos se superponen con regiones reguladoras anotadas experimentalmente.

La metáfora computacional ENCODE: los genes como rutinas "codificadas libremente"

La nueva perspectiva de ENCODE, por supuesto, no encaja con la metáfora del gen como una simple rutina invocable en un enorme sistema operativo. En esta nueva perspectiva, uno ingresa a una "rutina" de genes de muchas maneras diferentes en el marco de empalmes alternativos y entramados de transcripciones largas. La ejecución del sistema operativo genómico no tiene una calidad tan clara como esta idea de repetición.

llamadas tivas a una subrutina discreta en un sistema operativo de computadora normal. Sin embargo, el marco de descripción del genoma como código ejecutado todavía tiene algún mérito. Es decir, todavía se puede entender la transcripción de genes en términos de hilos paralelos de ejecución, con la advertencia de que estos hilos no siguen una estructura de subrutina modular canónica. Más bien, los hilos de ejecución se entrelazan de una manera bastante "sin orden ni concierto", muy parecido a lo que se describiría como un código de programa de computadora descuidado y sin estructura con muchas declaraciones GOTO entrando y saliendo de bucles y otras construcciones.

La importancia de los modelos genéticos para interpretar el experimento de alto rendimiento en ENCODE

Dados los provocativos hallazgos del proyecto ENCODE, uno se pregunta hasta qué punto se puede impulsar la interpretación de los experimentos de alto rendimiento. Esta interpretación es, de hecho, muy supeditada al uso de modelos genéticos.

Aspectos de la interpretación de datos de matrices de mosaico

Una gran parte de los datos de transcripción se generó utilizando micromatrices de mosaico de alta densidad (Emanuelsson et al. 2007; Rozowsky et al. 2007; The ENCODE Project Consortium 2007). La ventaja de tales arreglos es que prueban la transcripción de una manera imparcial y detallada, sin ideas preconcebidas sobre dónde buscar actividad. Por otro lado, el resultado de un experimento de matriz de mosaico puede ser ruidoso y necesita una interpretación cuidadosa para permitir la recopilación de un conjunto confiable de regiones transcritas. La cantidad de transcripción detectada depende en gran medida de los umbrales utilizados al llamar a las regiones transcritas y, en cierta medida, también de los algoritmos de segmentación utilizados para delimitar las regiones transcritas de las regiones no transcritas. Es más,

El resultado esperado exacto de un experimento de mapeo de transcripción, el verdadero mapa de transcripción, es, por supuesto, desconocido. Por lo tanto, una parte crucial de la interpretación de datos de matriz de mosaico de mapeo de transcripción es comprender cómo la señal es diferente de varias expectativas aleatorias (modelos nulos). Una forma ingenua de lograr este objetivo es aleatorizar los datos sin procesar y luego aplicar todos los esquemas de normalización, puntuación y segmentación (con parámetros sin cambios) para obtener una "línea de base" de transcripción que se obtiene utilizando datos supuestamente sin sentido. Pero no está claro si esta es la mejor manera de hacerlo: el contenido de GC, la distribución de longitud de las regiones transcritas (reales o esperadas), la composición de dinucleótidos y otras características también deben tenerse en cuenta en la distribución de transcripción "de referencia".

Los genes como modelos estadísticos que resumen muchos experimentos

En el contexto de la interpretación de experimentos de alto rendimiento, como matrices de mosaico, el concepto de gen tiene una importancia práctica adicional, como modelo estadístico para ayudar a interpretar y proporcionar un resumen conciso de datos experimentales potencialmente ruidosos.

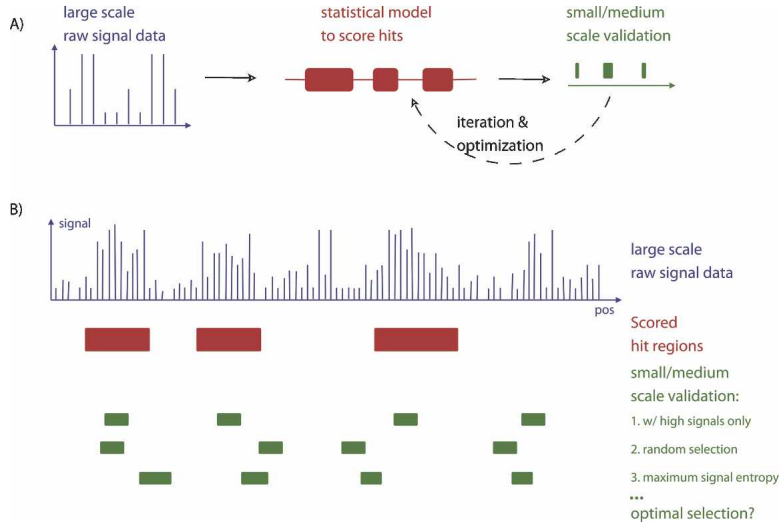

Figura 3. Formación de modelos de genes estadísticos basados en alta densidad microarrays de oligonucleótidos embaldosado datos. (*UN*) Los datos de señales a gran escala de los experimentos con matrices de mosaico se pueden usar para entrenar modelos estadísticos para calificar los aciertos, y una proporción pequeña o mediana de estos resultados se puede validar aún más mediante experimentos u otros conocimientos biológicos a través de iteraciones y optimizaciones. (*B*) Se pueden usar diferentes estrategias para seleccionar regiones genómicas para la validación; por ejemplo, (1) seleccionar solo las regiones con señales altas, (2) seleccionar regiones al azar, o (3) seleccionar aquellas que tienen las máximas entropías de señal, que generalmente contienen "bordes" de señales altas y bajas. Una pregunta que vale la pena hacer es si existe una forma óptima de selección que ayude mejor a entrenar el modelo estadístico.

Por ejemplo, los experimentos de matriz de mosaico transcripcional finalmente identifican TAR / transfrags, que generalmente corresponden a los exones en los genes. Por lo tanto, los modelos de genes más apropiados a considerar pueden ser gráficos de empalme (Heber et al. 2002) que representan exones como nodos y eventos de empalme como bordes dirigidos.

Para construir y ajustar modelos estadísticos para la interpretación experimental, es necesario combinar otros conocimientos biológicos relacionados (p. ej., anotación de genes y datos de validación experimental) con los datos de matriz de alto rendimiento. Por ejemplo, la matriz de datos de la transcripción puede identificar aislado regiones transcritas, y validación experimental como la raza puede proporcionar información de conectividad. Usando estos datos juntos, los modelos estadísticos pueden entrenarse mejor y luego pueden usarse para analizar el resto de los datos de alto rendimiento que no están cubiertos por los experimentos de validación.

Diferentes modelos estadísticos (Karplus et al. 1999; Bertone et al. 2004; Schadt et al. 2004; Gibbons et al. 2005; Ji and Wong 2005; Li et al. 2005; Du et al. 2006; Marioni et al. 2006) se han propuesto para explicar la generación de los datos de la matriz de mosaico. Como se muestra en la Figura 3, estos modelos pueden entrenarse utilizando los datos de la matriz de mosaico y otros conocimientos biológicos y luego extrapolarse a la secuencia del genoma completo para segmentarlo mejor en elementos funcionales. A medida que se acumula más y más conocimiento biológico, especialmente a través de la validación experimental de regiones funcionales predichas generadas por el procedimiento de análisis, podemos esperar que los modelos estén mejor entrenados, lo que conducirá a resultados de análisis refinados de estos experimentos. Sin embargo, será poco práctico validar cada uno de los elementos funcionales identificados mediante experimentos de mosaico utilizando RT-PCR o RACE. Para cada experimento de matriz de mosaico, quizás solo se valide experimentalmente un conjunto de tamaño mediano de regiones funcionales predichas.

Como se muestra en la Figura 3, las regiones para la validación experimental se pueden seleccionar utilizando diferentes estrategias. Obviamente, es beneficioso elegir estas regiones de una manera óptima para que el modelo entrenado en base a estos resultados de validación pueda analizar con mayor precisión. lyze el resto de los datos de la matriz de mosaico. En un caso específico, cuando se analizan datos de matrices de mosaico utilizando un modelo oculto de Markov (Du et al. 2006), si las regiones de validación se seleccionan para lograr la máxima entropía de señal, el esquema de selección MaxEntropy, el modelo de segmentación de genes resultante supera a otros. Para las matrices de mosaico transcripcional, MaxEntropy generalmente seleccionará regiones que contengan tanto exones como intrones.

Hacia una definición actualizada de un gen

Como hemos descrito anteriormente, nuestro conocimiento de los genes ha evolucionado mucho durante el último siglo. Si bien nuestra comprensión ha aumentado, también hemos descubierto un número cada vez mayor de aspectos problemáticos con definiciones simples de un gen (Tabla 1). El empalme (incluido el empalme alternativo) y la transcripción intergénica son obviamente algunos de los aspectos más problemáticos. Como se muestra en la Figura 4, la frecuencia de mención de estos términos en la literatura biológica ha ido aumentando considerablemente. Así, se preparó el escenario para el proyecto ENCODE y la gran complejidad en el aparato transcripcional y regulador que destacó. En este punto, no está claro qué hacer: en el extremo, podríamos declarar muerto el concepto del gen e intentar encontrar algo completamente nuevo que se ajuste a todos los datos. Sin embargo, sería difícil hacer esto con consistencia. Aquí, hicimos un intento tentativo de compromiso, diseñando actualizaciones y parches para la definición existente de un gen.

Criterios a considerar en la actualización de la definición

En primer lugar, consideramos que varios criterios son importantes al generar una definición actualizada para un gen: (1) Una nueva definición debe intentar ser *compatible con versiones anteriores*, en el sentido de que algo que solía llamarse gen debería seguir siendo un gen. (2) Debe ser *independiente del organismo*; es decir, ser tan válida para una bacteria como para un virus o un eucariota superior. (3) Debe ser una declaración de un *idea sencilla*, en lugar de enumerar varios mecanismos y excepciones. (4) Debe ser lo suficientemente práctico para que uno pueda *enumerar fácilmente los genes* y responder a una pregunta como "¿Cuántos genes hay en el genoma humano?" (5) Debería ser *compatible con otra nomenclatura biológica* que hace uso de la idea de un gen digital. Por ejemplo, debe ser coherente con el término reguloma, que representa el conjunto completo de interacciones reguladoras en un organismo.

Una definición actualizada propuesta

Hay tres aspectos de la definición que enumeraremos a continuación, antes de proporcionar la definición sucinta:

1. Un gen es una secuencia genómica (ADN o ARN) que codifica directamente moléculas producto funcionales, ya sea ARN o proteína.
2. En el caso de que existan varios productos funcionales que comparten regiones superpuestas, se toma elUnión de todas las secuencias genómicas superpuestas que las codifican.

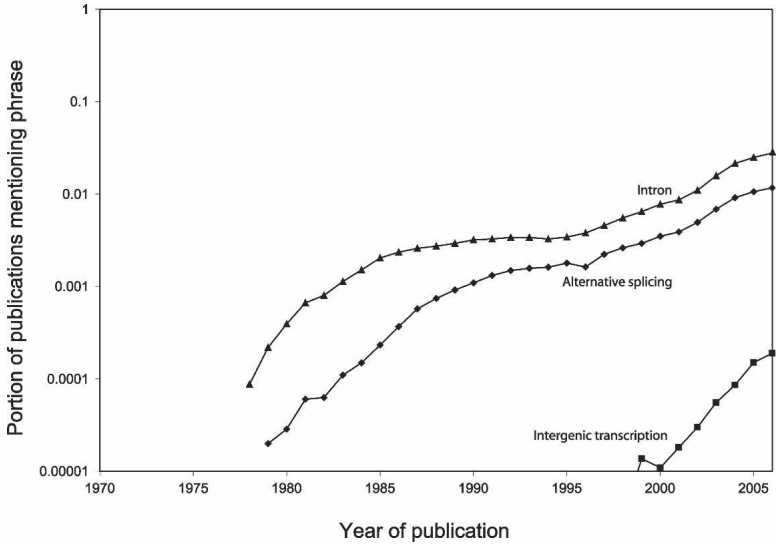

Figura 4. Análisis de palabras clave y complejidad de genes. Usando Google Scholar, una búsqueda de texto completo de Se realizaron artículos científicos para las palabras clave "intrón", "empalme alternativo" y "transcripción intergénica". Las pendientes de las curvas indican que en los últimos años ha aumentado la frecuencia de mención de términos relacionados con la complejidad de un gen. (La búsqueda de Google Scholar se limitó a artículos en las siguientes áreas temáticas: "Biología, Ciencias de la Vida y Ciencias Ambientales", "Química y Ciencias de los Materiales", "Medicina, Farmacología y Ciencias Veterinarias").

3. Esta unión debe sercoherente —es decir, hecho por separado para la proteína final y los productos de ARN— pero no requiere que todos los productos compartan necesariamente una subsecuencia común.

Esto se puede resumir de manera concisa como:

El gen es una unión de secuencias genómicas que codifican un conjunto de productos funcionales potencialmente superpuestos.

La figura 5 proporciona un ejemplo para ilustrar la aplicación de esta definición.

Aspectos e implicaciones de la definición

Hay implicaciones importantes de esta definición.

Colapso en casos simples

En casos simples donde el gen no es discontinuo o no hay productos superpuestos, nuestra definición colapsa a la versión clásica de ser una secuencia de ADN que codifica una proteína o un producto de ARN.

Proyectando hacia abajo en la unión

En nuestra definición propuesta de un gen, diferentes productos funcionales de la misma clase (proteína o ARN) que se superponen en su uso de la secuencia primaria de ADN se combinan en el mismo gen. Esta superposición se realiza proyectando la secuencia del producto final (ya sea secuencia de aminoácidos o de ARN) sobre la secuencia genómica original de la que se deriva. Uno podría, en principio, superponer las secuencias de los productos finales ("proyectar hacia arriba"); sin embargo, dado que la anotación de genes se realiza para el ADN genómico, creemos que nuestra elección es la más consistente con la práctica actual. Un punto obvio que aún debe señalarse es que, al observar productos genómicos con segmentos de secuencia comunes, la mera identidad de secuencia no es suficiente; los productos tienen que ser codificados directamente desde la misma región genómica. Por lo tanto, las proteínas parálogas pueden compartir bloques de secuencia, las secuencias que los codifican residen en lugares separados del genoma, por lo que no constituirían un solo gen.

Exones con marco desplazado

Hay casos, como el del gen supresor de tumores CDKN2A (anteriormente INK4a/ARF) (p. ej., Quelle et al. 1995), en los que un pre-ARNm puede empalmarse alternativamente para generar un ARNm con un cambio de marco en la secuencia de la proteína. Por lo tanto, aunque los dos ARNm tienen secuencias de codificación en común, los productos proteicos pueden ser completamente diferentes. Este caso bastante inusual plantea la cuestión de cómo se debe manejar exactamente la identidad de secuencia al tomar la unión de segmentos de secuencia que se comparten entre productos proteicos. Si uno considera la secuencia de los productos proteicos, hay dos proteínas no relacionadas, por lo que debe haber dos genes con conjuntos de secuencias superpuestas. Si uno "proyecta" la secuencia de los productos proteicos de regreso a la secuencia de ADN que los codificó (como se describió anteriormente), entonces hay dos conjuntos de secuencias con elementos comunes, por lo que hay un gen. El hecho de que las secuencias de estas dos proteínas estén restringidas simultáneamente, de modo que una mutación en una de ellas afectaría simultáneamente a la otra, sugiere que esta situación no es similar a la de dos genes codificadores de proteínas no relacionados. Por esta razón, generalizando a partir de este caso especial, favorecemos el método de tomar la unión de los segmentos de secuencia, no de los productos, sino de las secuencias de ADN que codifican para las secuencias de productos.

Regiones reguladoras no incluidas

Aunque las regiones reguladoras son importantes para la expresión génica, sugerimos que no deben considerarse al decidir si múltiples productos pertenecen al mismo gen. Este aspecto de la definición resulta de nuestro concepto del operón bacteriano. Tradicionalmente no se ha considerado que el hecho de que los genes en un operón compartan una región operadora y promotora implique que sus productos proteicos sean productos alternativos de un solo gen. En consecuencia, en los eucariotas superiores, dos transcritos que se originan en el mismo sitio de inicio de la transcripción (que comparten el mismo promotor y elementos reguladores) pero que no comparten ningún elemento de secuencia en sus productos finales (p. ej., debido a empalmes alternativos) no serían productos de la misma gen. Una lógica similar se aplicaría a múltiples transcripciones que comparten un potenciador o aislante común pero distante.

Productos finales, no grupos de transcripción

Como la definición actualizada enfatiza los productos finales de un gen, no tiene en cuenta los productos intermedios que se originan en una región genómica que puede superponerse. Por ejemplo, una transcripción intrónica claramente comparte secuencias con una transcripción más grande superpuesta, pero este hecho es irrelevante cuando concluimos que

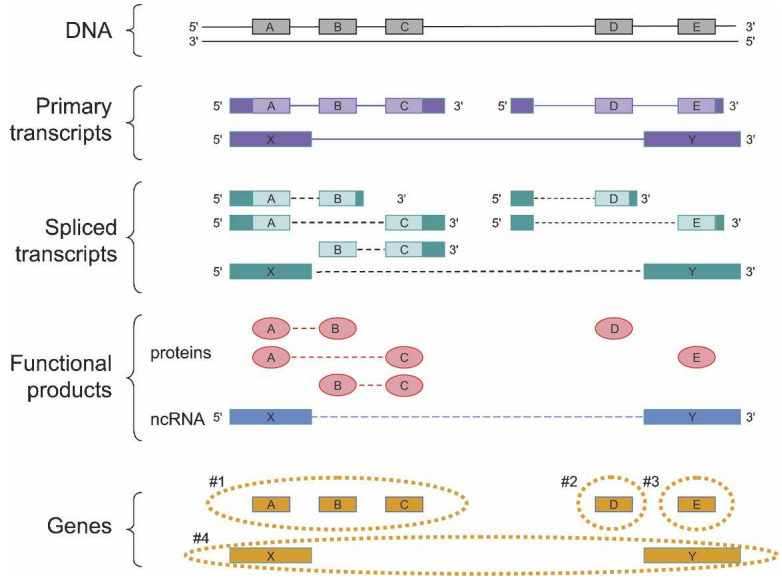

Figura 5. Cómo se puede aplicar la definición propuesta del gen a un caso de muestra. Una genómica región produce tres transcripciones primarias. Después del empalme alternativo, los productos de dos de estos codifican cinco productos proteicos, mientras que el tercero codifica un producto de ARN no codificante (ncRNA). Los productos proteicos están codificados por tres grupos de segmentos de secuencia de ADN (A, B y C; D y E). En el caso del grupo de tres segmentos (A, B, C), cada segmento de secuencia de ADN es compartido por al menos dos de los productos. Dos transcripciones primarias comparten una región no traducida de 5, pero sus regiones traducidas D y E no se superponen. También hay un producto de ARN no codificante, y debido a que su secuencia es de ARN, no de proteína, el hecho de que comparta sus secuencias genómicas (X e Y) con los segmentos genómicos codificadores de proteínas A y E no lo convierte en un coproducto. de estos genes codificadores de proteínas. En resumen, hay cuatro genes en esta región, y son los conjuntos de secuencias que se muestran dentro de las líneas discontinuas de color naranja: el gen 1 consta de los segmentos de secuencia A, B y C; el gen 2 consta de D; gen 3 de E; y el gen 4 de X e Y. En el diagrama, para mayor claridad, las secuencias exónicas y proteína A-E han sido alineados verticalmente, por lo que las líneas de trazos para las transcripciones empalmados y productos funcionales indican la conectividad entre las secuencias de proteínas (óvalos) y Secuencias de ARN (cajas). (Recuadros sólidos en las transcripciones) Secuencias no traducidas, (recuadros abiertos) secuencias traducidas. por lo que las líneas discontinuas para las transcripciones empalmadas y los productos funcionales indican la conectividad entre las secuencias de proteínas (óvalos) y las secuencias de ARN (recuadros). (Recuadros sólidos en las transcripciones) Secuencias no traducidas, (recuadros abiertos) secuencias traducidas. por lo que las líneas discontinuas para las transcripciones empalmadas y los productos funcionales indican la conectividad entre las secuencias de proteínas (óvalos) y las secuencias de ARN (recuadros). (Recuadros sólidos en las transcripciones) Secuencias no traducidas, (recuadros abiertos) secuencias traducidas.

los dos productos no comparten bloques de secuencia. Este concepto se puede generalizar a otros tipos de genes discontinuos, como los genes reorganizados (p. ej., en el locus del gen de la inmunoglobulina, el segmento C es común a todos los productos proteicos codificados a partir de él), o *trans*transcripciones -spliced (donde una pre-mRNA se puede empalmar a una serie de otros pre-mRNAs antes del procesamiento posterior y la traducción). Esto implica que el número de genes en el genoma humano va a aumentar significativamente cuando se haya completado la encuesta del transcriptoma humano. A la luz de la gran cantidad de transcripciones entrelazados que fueron identificados por el consorcio ENCODE, si tratamos a agruparse transcripciones enteras juntos para formar la superposición de grupos de transcripción (una definición alternativa potencial de un gen), entonces se verá que grandes segmentos de cromosomas haría fundirse en estos grupos. Esta definición alternativa de un gen resultaría en un número mucho menor "genes", y sería de utilidad limitada.

Splicing alternativo

En relación con los productos génicos empalmados alternativamente, existe la posibilidad de que ningún exón codificante se comparta entre todos los productos proteicos. En este caso, se entiende que la unión de estos segmentos de secuencia define el gen, siempre que cada exón sea compartido entre al menos dos miembros de este grupo de productos.

UTR

Las regiones no traducidas (UTR) 5 y 3 juegan un papel importante en la traducción, regulación, estabilidad y/o localización de los ARNm.

Cuando se usa una definición estricta de regiones que codifican el producto final de un gen codificador de proteínas, estas regiones ya no se considerarían parte del gen, como suele ser el caso en el uso actual. Además, los transcritos que codifican proteínas que comparten la secuencia de ADN solo en sus regiones no traducidas o intrones no se agruparían en un gen común. Al eliminar los UTR de la definición de un gen, se puede evitar el problema de múltiples extremos 5 y 3 que nublan la delineación del gen y también evitar una situación en la que aguas arriba o *trans*5- Las secuencias líder se empalman en una secuencia codificante de proteína. Además, se ha observado que la mayoría de las transcripciones de codificación de proteínas más largas identificadas por ENCODE difieren solo en sus UTR y, por lo tanto, nuestra definición es bastante transparente para este grado de complejidad de la transcripción.

Regiones asociadas a genes

Como se describió anteriormente, las regiones reguladoras y no traducidas que desempeñan un papel importante en la expresión génica ya no se considerarían parte del gen. Sin embargo, nos gustaría crear una "categoría" especial para ellos, diciendo que serían*asociado al gen*. De esta manera, estas regiones aún conservan su importante papel en la contribución a la función del gen. Además, su capacidad para controlar tributo a la expresión de varios genes puede ser reconocido. Esto es particularmente cierto para los elementos de largo alcance, como la betaglobina LCR, que contribuye a la expresión de varios genes, y probablemente será el caso de muchos otros potenciadores a medida que se mapean sus verdaderos objetivos genéticos. También se puede aplicar a regiones no traducidas que contribuyen a múltiples loci de genes, como las transcripciones empalmadas largas observadas en la región ENCODE y*trans*-exones empalmados.

Conjuntos inconexos de secuencia genómica

Para mayor claridad en la discusión, nos referimos a "ADN" cuando nos referimos a secuencias genómicas en general. Nuestra definición propuesta es aplicable a todos los genomas, incluido el de los virus de ARN. En casos complejos, resulta que el gen no corresponde a un solo locus genético discreto, ya que las secuencias que codifican sus productos pueden estar muy separadas en el genoma. En particular, debido a que el gen es un conjunto de secuencias compartidas entre los productos, no existe un requisito de conectividad entre estas secuencias y las secuencias que las conectan no necesitan ser parte del gen. Por lo tanto, los miembros de una secuencia pueden estar en diferentes hebras de un cromosoma o incluso en cromosomas separados. Esto significa que*trans*-las transcripciones empalmadas pertenecen a un gen.

Conclusión: ¿Qué es la función?

La visión clásica de un gen como una unidad de información hereditaria alineada a lo largo de un cromosoma, cada uno codificando para una proteína, tiene

cambió drásticamente durante el siglo pasado. Para Morgan, los genes en los cromosomas eran como cuentas en un hilo. La revolución de la biología molecular cambió considerablemente esta idea. Para citar a Falk (1986), ''. . . el gen es [. . .] ni discreto [. . .] ni continuo [. . .], ni tiene una ubicación constante [. . .], ni una función clara [. . .], ni siquiera secuencias constantes [. . .] ni límites definidos". Y ahora el proyecto ENCODE ha incrementado aún más la complejidad.

Lo que no ha cambiado es que el genotipo determina el fenotipo y, a nivel molecular, esto significa que las secuencias de ADN determinan las secuencias de las moléculas funcionales. En el caso más simple, una secuencia de ADN todavía codifica una proteína o ARN. Pero en el caso más general, podemos tener genes que consisten en módulos de secuencia que se combinan de múltiples formas para generar productos. Al centrarse en los productos funcionales del genoma, esta definición establece un estándar concreto al enumerar sin ambigüedades el número de genes que contiene.

Un aspecto importante de nuestra definición propuesta es el requisito de que la proteína o los productos de ARN deben ser *funcional* con el fin de asignarlos a un gen particular. Creemos que este se conecta al principio básico de la genética, que determina el genotipo fenotipo. A nivel molecular, se supone que el fenotipo se refiere a la función bioquímica. Nuestra intención es hacer que nuestra definición compatibles con versiones anteriores con los conceptos anteriores del gen.

Este énfasis en los productos funcionales, por supuesto, destaca la cuestión de qué es realmente la función biológica. Con esto, pasamos a la difícil pregunta de "¿qué es un gen?" a "¿qué es una función?"

Se necesitarán ensayos bioquímicos y mutacionales de alto rendimiento para definir la función a gran escala (Lan et al. 2002, 2003). Con suerte, en la mayoría de los casos será solo cuestión de tiempo hasta que adquiramos la evidencia experimental que establecerá lo que hacen la mayoría de los ARN o proteínas. Hasta entonces, tendremos que usar términos de "marcador de posición" como TAR, o indicar nuestro grado de confianza para asumir la función de un producto genómico. También podemos inferir la funcionalidad de las propiedades estadísticas de la secuencia (p. ej., Ponjavic et al. 2007).

Sin embargo, probablemente nunca seremos capaces de conocer la función de todas las moléculas en el genoma. Es concebible que algunos productos genómicos sean simplemente "ruido", es decir, resultados de eventos evolutivamente neutrales que son tolerados por el organismo (p. ej., Tress et al. 2007). O bien, puede haber una función compartida por tantos otros productos genómicos que identificar la función mediante enfoques mutacionales puede ser muy difícil. Si bien determinar la función biológica puede ser difícil, probar la falta de función es aún más difícil (casi imposible). Es probable que algunos bloques de secuencia en el genoma mantengan sus etiquetas de "TAR de función desconocida" indefinidamente. Si tales regiones comparten secuencias con genes funcionales, sus límites (o más bien, la pertenencia a su conjunto de secuencias) seguirán siendo inciertos.

Expresiones de gratitud

Agradecemos al consorcio ENCODE y reconocemos las siguientes fuentes de financiación: subvención ENCODE #U01HG03156 del Instituto Nacional de Investigación del Genoma Humano (NHGRI)/ Institutos Nacionales de Salud (NIH); NIH Grant T15 LM07056 de la Biblioteca Nacional de Medicina (CB, ZDZ); y una beca internacional saliente Marie Curie (JOK).

Referencias

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. y Sorek, R. 2006. Fusión de genes mediada por transcripción en el genoma humano. *Genoma Res.* **dieciséis:** 30–36.

Avery, OT, MacLeod, CM y McCarty, M. 1944. Estudios sobre la naturaleza química de la sustancia que induce la transformación de los tipos neumocócicos. *Exp. J. Medicina.* **79:** 137–158.

Balakirev, ES y Ayala, FJ 2003. Pseudogenes: Are they "junk" or ¿ADN funcional? *año Rev. Genet.* **37:** 123–151.

Beadle, GW y Tatum, EL 1941. Control genético de bioquímica reacciones en Neurospora. *proc. nacional Academia ciencia* **27:** 499–506. Benzer, S. 1955. Estructura fina de una región genética en bacteriófagos. *proc. nacional Academia ciencia* **41:** 344–354.

Berget, SM, Moore, C. y Sharp, PA 1977. Segmentos empalmados en el 5-terminal del ARNm tardío del adenovirus 2. *proc. nacional Academia ciencia* **74:** 3171–3175.

Bertone, P., Stolc, V., Royce, TE, Rozowsky, JS, Urban, AE, Zhu, X., Rinn, JL, Tongprasit, W., Samanta, M., Weissman, S., et al. 2004. Identificación global de secuencias transcritas humanas con matrices de mosaico del genoma. *Ciencia* **306:** 2242–2246.

Blumenthal, T. 2005. Empalme trans y operones. WormBook (ed. El *C. elegans* comunidad de investigación). WormBook, doi/ 10.1895/ wormbook.1.5.1, http://www.wormbook.org.

Borst, P. 1986. Transcripción discontinua y variación antigénica en tripanosomas. *año Rev. Bioquímica.* **55:** 701–732.

Cawley, S., Bekiranov, S., Ng, HH, Kapranov, P., Sekinger, EA, Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, AJ, et al. 2004. El mapeo imparcial de los sitios de unión del factor de transcripción a lo largo de los cromosomas humanos 21 y 22 apunta a una regulación generalizada de los ARN no codificantes. *Celúla* **116:** 499–509.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Mapas transcripcionales de 10 cromosomas humanos con una resolución de 5 nucleótidos. *Ciencia* **308:** 1149–1154.

Chow, LT, Gelinas, RE, Broker, TR y Roberts, RJ 1977. Un asombrosa disposición de la secuencia en los extremos 5 del ARN mensajero del adenovirus 2. *Celúla* **12:** 1–8.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P. y Duret, L. 2002. Análisis comparativo de secuencias de la región del centro de inactivación X en ratones, humanos y bovinos. *Genoma Res.* **12:** 894–908.

Contreras, R., Rogiers, R., Van de, VA y Fiers, W. 1977. Superposición del gen VP2-VP3 y del gen VP1 en el genoma SV40. *Celúla* **12:** 529–538.

Crick, FHC 1958. Sobre la síntesis de proteínas. *Síntoma Soc. Exp. Biol.* **XII:** 138–163.

Dawkins, R. 1976. *El gen egoísta*. Prensa de la Universidad de Oxford, Oxford, Reino Unido. Denoeud, F., Kapranov, P., Ucla, C., Frankish, A., Castelo, R., Drenkow, J., Lagarde, J., Alioto, T., Manzano, C., Chrast, J., et al. 2007. Uso destacado de los sitios de inicio de la transcripción distal 5 y descubrimiento de una gran cantidad de exones adicionales en las regiones ENCODE. *Genoma Res.* (este problema) doi: 10.1101/gr5660607.

Dobrovic, A., Gareau, JL, Ouellette, G. y Bradley, WE 1988. ADN metilación e inactivación genética en el locus de la timidina quinasa: dos mecanismos diferentes para silenciar genes autosómicos. *Somat. Mol celular. Gineta.* **14:** 55–68.

Doolittle, R. 1986. *De URF y ORF: una introducción sobre cómo analizar derivados secuencias de aminoácidos*. Libros de ciencia universitaria, Mill Valley, CA.

Du, J., Rozowsky, JS, Korbel, J., Zhang, ZD, Royce, TE, Schultz, MH, Snyder, M. y Gerstein, M. 2006. Un marco de modelo de Markov oculto supervisado para segmentar eficientemente datos de matriz de mosaico en experimentos transcripcionales y de chip-chip: incorporar sistemáticamente conocimiento biológico validado. *Bioinformática* **22:** 3016–3024.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. y Avner, P. 2006. El gen de ARN Xist evolucionó en euterios por pseudogenización de un gen codificador de proteínas. *Ciencia* **312:** 1653–1655.

Early, P., Huang, H., Davis, M., Calame, K. y Hood, L. 1980. Un El gen de la región variable de la cadena pesada de la inmunoglobulina se genera a partir de tres segmentos de ADN: VH, D y JH. *Celúla* **19:** 981-992. Eddy, SR 2001. No-codificación de los genes de ARN y ARN del mundo moderno. *Nat. Rev. Genet.* **2:** 919–929.

Eisen, H. 1988. Edición de ARN: ¿Quién está primero? *Celúla* **53:** 331-332.

Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, JS, Urban, AE, Du, J., Lian, Z., Stolc, V., Weissman, S., Snyder, M., et al. 2007. Evaluación del rendimiento de diferentes estrategias de microarrays de mosaico de alta densidad para mapear regiones transcritas del genoma humano. *Genoma Res.* (este número) doi: 10.1101/gr.5014606.

El Consorcio del Proyecto ENCODE. 2007. Identificación y análisis de elementos funcionales en el 1% del genoma humano por la ENCODE

proyecto piloto.*Naturaleza*(en prensa).
Euskirchen, G., Royce, TE, Bertone, P., Martone, R., Rinn, JL, Nelson, FK, Sayward, F., Luscombe, NM, Miller, P., Gerstein, M., et al. 2004. CREB se une a múltiples loci en el cromosoma 22 humano.*mol. Celúla. Biol.***24:** 3804–3814.
Falk, R. 1986. ¿Qué es un gen?*Semental. hist. Filosofía ciencia***17:**133–173.
Fiers, W., Contreras, R., De Wachter, R., Haegeman, G., Merregaert, J., Jou, WM y Vandenberghe, A. 1971. Avances recientes en la determinación de la secuencia del ARN del bacteriófago MS2. *bioquímica* **53:**495–506.
Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., MinJou, W., Molemans, F., Raeymakers, A., Van den Berghe, A., et al. 1976. Secuencia de nucleótidos completa del ARN del bacteriófago MS2: estructura primaria y secundaria del gen de la replicasa.*Naturaleza***260:**500–507.
Fleischmann, RD, Adams, MD, White, O., Clayton, RA, Kirkness, EF, Kerlavage, AR, Construido, CJ, Tumba, JF, Dougherty, BA, Merrick, JM, et al. 1995. Secuenciación aleatoria del genoma completo y montaje de*Haemophilus influenzae*Calle*Ciencia***269:**496–512.
Frith, MC, Wilming, LG, Forrest, A., Kawaji, H., Tan, SL, Wahlestedt, C., Bajic, VB, Kai, C., Kawai, J., Carninci, P., et al. 2006. ARN pseudo-mensajero: Fantasmas del transcriptoma.*PLoS Genet.* **2:**e23.

Gelinas, RE y Roberts, RJ 1977. Una predominante 5--undecanucleótido en ARN mensajero tardío de adenovirus 2.*Celúla* **11:** 533–544.
Gibbons, FD, Proft, M., Struhl, K. y Roth, FP 2005. Astilladora: Descubrimiento de objetivos de factores de transcripción a partir de micromatrices de inmunoprecipitación de cromatina mediante estabilización de la varianza. *Genoma Biol.* **6:**R96.
Gingeras, T. 2007. Origen de los fenotipos: Genes y transcripciones.*genoma Res.*(este número) doi: 10.1101/gr.625007.
Griffith, F. 1928. La importancia de los tipos neumocócicos.*J. hig. (Londres)* **27:**113–159.
Griffiths, PE y Stotz, K. 2006. Genes in the postgenomic era.*teor. Medicina. Bioeth.***27:**499–521.
Handa, H., Bonnard, G. y Grienenberger, JM 1996. La colza El gen mitocondrial que codifica un homólogo de la proteína bacteriana Ccl1 se divide en dos marcos de lectura transcritos de forma independiente. *mol. Gen. Genet.***252:**293–302.
Harrison, PM, Zheng, D., Zhang, Z., Carriero, N. y Gerstein, M. 2005. Pseudogenes procesados transcritos en el genoma humano: una forma intermedia de retrosecuencia expresada que carece de capacidad de codificación de proteínas.*Ácidos Nucleicos Res.***33:**2374–2383.
Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, CK, Chrast, J., Lagarde, J., Gilbert, JG, Storey, R., Swarbreck, D., et al. 2006. GENCODE: Producción de una anotación de referencia para ENCODE.*Genoma Biol.***7 suplemento 1:**S4.1–S9.
Heber, S., Alekseyev, M., Sze, S., Tang, H. y Pevzner, PA 2002. Gráficos de empalme y problema de ensamblaje de EST.*Bioinformática* **18:**S181-S188.
Heimans, J. 1962. Hugo de Vries y el concepto de gen.*Soy. Nat.* **96:**93–104.
Henikoff, S., Keene, MA, Fechtel, K. y Fristrom, JW 1986. Gene dentro de un gen: Anidado*drosófila*los genes codifican proteínas no relacionadas en hebras opuestas de ADN.*Celúla***44:**33–42.
Hershey, AD y Chase, M. 1955. Un límite superior a la proteína contenido de la sustancia germinal del bacteriófago T2.*Virología* **1:** 108–127.
Iafrate, AJ, Feuk, L., Rivera, MN, Listewnik, ML, Donahoe, PK, Qi, Y., Scherer, SW y Lee, C. 2004. Detección de variación a gran escala en el genoma humano.*Nat. Gineta.***36:**949–951.
Jacob, F. y Monod, J. 1961. Mecanismos reguladores genéticos en el síntesis de proteínas.*J. Mol. Biol.***3:**318–356.
Ji, H. y Wong, WH 2005. TileMap: Create chromosomal map of hibridaciones de matriz de mosaico.*Bioinformática***21:**3629–3636.
Johannsen, W. 1909. Elemente der exakten Erblichkeitslehre, Jena. Citado por Nils Roll-Hansen (1989). El experimento crucial de Wilhelm Johannsen.*Biol. Filosofía***4:**303–329.
Kapranov, P., Cawley, SE, Drenkow, J., Bekiranov, S., Strausber, RL, Fodor, SP, y Gingeras, TR actividad transcripcional de 2002. gran escala en los cromosomas 21 y 22.*Ciencia***296:**916–919. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. y Hughey, R. 1999. Predicción de la estructura de la proteína usando solo información de secuencia.*Proteínas***37 (suplemento 3):**121–125.
Kim, TH, Barrera, LO, Zheng, M., Qu, C., Singer, MA, Richmond, TA, Wu, Y., Green, RD y Ren, B. 2005. Un mapa de alta resolución de promotores activos en el genoma humano.*Naturaleza***436:**876–880.
Korneev, SA, Park, JH y O'Shea, M. 1999. Expresión neuronal de La proteína neural óxido nítrico sintasa (nNOS) es suprimida por un ARN antisentido transcrito a partir de un pseudogen NOS.*J. Neurosci.* **19:**7711–7720.
Lan, N., Jansen, R. y Gerstein, M. 2002. Hacia una sistemática definición de la función de la proteína que escala al nivel del genoma: definición de la función en términos de interacciones.*proc. IEEE***90:**1848–1858. Lan, N., Montelione, GT y Gerstein, M. 2003. Ontologías para Proteómica: Hacia una definición sistemática de estructura y función que escala al nivel del genoma.*actual Opinión química Biol.* **7:**44–54.

Lander, ES, Linton, LM, Birren, B., Nusbaum, C., Zody, MC, Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Secuenciación inicial y análisis del genoma humano.*Naturaleza* **409:**860–921.
Li, W., Meyer, CA y Liu, XS 2005. Un modelo oculto de Markov para análisis de experimentos ChIP-chip en matrices de mosaico del genoma y su aplicación a secuencias de unión a p53.*Bioinformática***21:**i274–i282.
Lindblad-Toh, KCM, Wade, TS, Mikkelsen, EK, Karlsson, DB, Jaffe, M., Kamal, M., Abrazadera, JL, Chang, EJ, Kulbokas 3rd, MC, Zody, E., et al. 2005. secuencia del genoma, el análisis comparativo y la estructura de haplotipos del perro doméstico.*Naturaleza***438:**803–819.
Lodish, H., Scott, MP, Matsudaira, P., Darnell, J., Zipursky, L., Kaiser, CA, Berk, A. y Krieger, M. 2000.*biología celular molecular*, 5ª ed. Freeman and Co., Nueva York.
Marioni, JC, Thorne, NP y Tavare, S. 2006. BioHMM: A modelo oculto heterogéneo de Markov para segmentar datos CGH de matriz.*Bioinformática***22:**1144-1146.
Mattick, JS y Makunin, IV 2006. ARN no codificante.*Tararear. mol. Gineta.***15 espec. Nº 1:**R17–R29.
McClintock, B. 1929. Un estudio citológico y genético del maíz triploide. *Genética***14:**180–222.
McClintock, B. 1948. loci mutable en el maíz.*Instituto Carnegie Lavado Año Libro***47:**155–169.
Mendel, JG 1866. Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn 4 Abhandlungen, 3–47. Citado por Robert C. Olby (1997) en http://www.mendelweb.org/MWolby.html, consultado el 16 de marzo de 2007. Morgan, TH, Sturtevant, AH, Muller, HJ y Bridges, CB 1915.*Él mecanismo de la herencia mendeliana.*Holt Rinehart & Winston, Nueva York.
Muller, HJ 1927. Transmutación artificial del gen.*Ciencia***46:**84–87. Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. y O'Neal, C. 1965. Códigos de RNA y síntesis de proteínas, VII. Sobre la naturaleza general del código de ARN.*proc. nacional Academia ciencia***53:**1161–1168.
Ohno, S. 1972. Tanto ADN "basura" en el genoma. En*Evolución de sistemas genéticos*, vol. 23 (ed. HH Smith), págs. 366–370. Simposios de Brookhaven en Biología. Gordon & Breach, Nueva York.
Parra, G., Reymond, A., Dabbouseh, N., Dermitzakis, ET, Castelo, R., Thomson, TM, Antonarakis, SE y Guigó, R. 2006. El quimerismo en tándem como medio para aumentar la complejidad de las proteínas en el genoma humano.*Genoma Res.***dieciséis:**37–44.
Paul, J. 1972. Teoría general de la estructura cromosómica y el gen. activación en eucariotas.*Naturaleza***238:**444–446.
Pearson, H. 2006. Genética: ¿Qué es un gen?*Naturaleza***441:**398–401.
Pedersen, JS, Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, ES, Kent, J., Miller, W. y Haussler, D. 2006. Identificación y clasificación de estructuras secundarias de ARN conservadas en el genoma humano.*Cómputo PLoS. Biol.*doi: 10.1371/journal.pcbi.0020033.
Ponjavic, J., Ponting, CP y Lunter, G. 2007. Funcionalidad o ruido transcripcional? Evidencia de selección dentro de ARN largos no codificantes.*Genoma Res.***17:**556–565.
Quelle, DE, Zindy, F., Ashmun, RA y Sherr, CJ 1995. Alternativa Los marcos de lectura del gen supresor de tumores INK4a codifican dos proteínas no relacionadas capaces de inducir la detención del ciclo celular.*Celúla* **83:**993–1000.
Rheinberger, HG 1995. ¿Cuándo leyó Darl Correns el libro de Gregor Mendel? ¿papel?*isis***86:**612–616.
Rinn, JL, Euskirchen, G., Bertone, P., Martone, R., Luscombe, NM, Hartman, S., Harrison, PM, Nelson, FK, Mille, P., Gerstein, M., et al. 2003. La actividad transcripcional del cromosoma 22 humano. *Genes y desarrollo***17:**529–540.
Rogic, S., Mackworth, AK y Ouellette, FB 2001. Evaluación de programas de secuencias de mamíferos gen de investigación.*Genoma Res.* **11:** 817–832.
Rozowsky, J., Newburger, D., Sayward, F., Wu, J., Jordan, G., Korbel, JO, Nagalakshmi, U., Yang, J., Zheng, D., Guigo, R., et al. 2007. La clasificación DART de la transcripción no anotada dentro de las regiones ENCODE: Asociación de la transcripción con loci conocidos y novedosos. *Genoma Res.*(este número) doi: 10.1101/gr.5696007.
Sager, R. y Kitchin, R. 1975. Silenciamiento selectivo del ADN eucariótico. *Ciencia***189:**426–433.

Schadt, EE, Edwards, SW, Guhathakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, KW, Russell, A., Li, G., et al. 2004. Un índice de transcripción completo del genoma humano generado mediante micromatrices y enfoques computacionales. *Genoma Biol.* **5:**R73.

Searls, DB 1997. Resumen: Enfoques lingüísticos de la biología secuencias. *computar aplicación Biosci.* **13:**333–344.

Searls, DB 2001. Lectura del libro de la vida. *Bioinformática* **17:**579–580.

Searls, DB 2002. El lenguaje de los genes. *Naturaleza* **420:**211-217. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., et al. 2004. Polimorfismo del número de copias a gran escala en el genoma humano. *Ciencia* **305:** 525–528.

Shi, Y., Seto, E., Chang, LS y Shen, KT 1991. Transcripcional represión por YY1, una proteína relacionada con GLI-Kruppel humana, y alivio de la represión por la proteína E1A del adenovirus. *Celúla* **67:**377–388.

Söll, D., Ohtsuka, E., Jones, DS, Lohrmann, R., Hayatsu, H., Nishimura, S. y Khorana, HG 1965. Estudios sobre polinucleótidos, XLIX. Estimulación de la unión de aminoacil-sRNA a ribosomas por ribotrinucleótidos y estudio de asignaciones de codones para 20 aminoácidos. *proc. nacional Academia ciencia* **54:**1378-1385.

Spilianakis, C., Lalioti, M., Town, T., Lee, G. y Flavell, R. 2005. Asociaciones intercromosómicas entre loci expresados alternativamente. *Naturaleza* **435:**637–645.

Sturtevant, H. 1913. La disposición lineal de seis factores ligados al sexo en *drosófila* como lo muestra su modo de asociación. *Exp. J. Zool.* **14:** 43–59.

Takahara, T., Kanazu, SI, Yanagisawa, S. y Akanuma, H. 2000. Los ARNm heterogéneos de Sp1 en células HepG2 humanas incluyen un producto de homotípico *trans*-empalme. *J. Biol. química* **275:**38067–38072.

Torrents, D., Suyama, M., Zdobnov, E. y Bork, P. 2003. A Estudio de todo el genoma de los pseudogenes humanos. *Genoma Res.* **13:**2559–2567.

Tress, M., Martelli, PL, Frankish, A., Reeves, G., Wesselink, JJ, Yeats, C., Olason, PI, Albrecht, M., Hegyi, H., Giorgetti, A., et al. 2007. Las implicaciones del empalme alternativo en el complemento proteico ENCODE. *proc. nacional Academia ciencia* **104:**5495–5500. Tschermak, E. 1900. Über Künstliche Kreuzung bei Pisum sativum. *Berichte Deutsche Botanischen. Gesellschaft* **18:**232–239. Tuan, DY, Solomon, WB, London, IM y Lee, DP 1989. Una potenciador independiente de la etapa de desarrollo, específico de eritroides, muy por encima de los genes humanos "--like globin". *proc. nacional Academia ciencia* **86:** 2554–2558.

Tuzun, E., Sharp, AJ, Bailey, JA, Kaul, R., Morrison, VA, Pertz, LM, Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. 2005. Variación estructural a escala fina del genoma humano. *Nat. Gineta.* **37:**727–732.

Vanin, EF, Goldberg, GI, Tucker, PW y Smithies, O. 1980. A ratón: pseudogén relacionado con la globina que carece de secuencias intermedias. *Naturaleza* **286:**222–226.

Venter, JC, Adams, MD, Myers, EW, Li, PW, Mural, RJ, Sutton, GG, Smith, HO, Yandell, M., Evans, CA, Holt, RA, et al. 2001. La secuencia del genoma humano. *Ciencia* **291:**1304–1351. Villa-Komaroff, L., Guttman, N., Baltimore, D. y Lodishi, HF 1975. Traducción completa del ARN del poliovirus en un sistema libre de células eucariotas. *proc. nacional Academia ciencia* **72:**4157–4161.

Vries, H. 1900. Sur la loi de disjonction des hybrides. *Comptes rendus de l'Académie des Sciences (París).* **130:**845–847.

Wade, N. 2003. *Termina el sorteo de Gene, pero el ganador bien puede estar equivocado. Nuevo tiempos de york*. http://query.nytimes.com/gst/fullpage.html?sec=health&res=9A02E0D81230F930A35755C0A9659C8B63 Wain, HM, Bruford, EA, Lovering, RC, Lush, MJ, Wright, MW, y Povey, S. 2002. Directrices para la nomenclatura de genes humanos. *genómica* **79:**464–470.

Washietl, S., Hofacker, IL, Lukasser, M., Huttenhofer, A. y Stadler, PF 2005. El mapeo de estructuras secundarias de ARN conservadas predice miles de ARN no codificantes funcionales en el genoma humano. *Nat. Biotecnología* **23:**1383–1390.

Washietl, S., Pedersen, JS, Korbel, JO, Stocsits, C., Gruber, AR, Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., et al. 2007. ARN estructurados en las regiones seleccionadas ENCODE del genoma humano. *Genoma Res.* (este número) doi: 10.1101/gr.5650707.

Waterston, RH, Lindblad-Toh, K., Birney, E., Rogers, J., Abril, JF, Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Secuenciación inicial y análisis comparativo del genoma del ratón. *Naturaleza* **420:**520–562.

Watson, JD y Crick, FHC 1953. Una estructura de desoxirribonucleico ácido. *Naturaleza* **171:**964-967.

Wold, F. 1981. Modificación química in vivo de proteínas (modificación post-traduccional). *año Rev. Bioquímica.* **50:**783–814. Yano, Y., Saito, R., Yoshida, N., Yoshiki, A., Wynshaw-Boris, A., Tomita, M. y Hirotsune, S. 2004. Un nuevo papel para los pseudogenes expresados como ncRNA: Regulación de la estabilidad del mRNA de su gen codificante homólogo. *J. Mol. Medicina.* **82:**414–422.

Zhang, Z., Harrison, PM, Liu, Y. y Gerstein, M. 2003. Millones de años de evolución conservados: un catálogo completo de los pseudogenes procesados en el genoma humano. *Genoma Res.* **13:**2541–2558.

Zhang, ZD, Paccanaro, A., Fu, Y., Weissman, S., Weng, Z., Chang, J., Snyder, M. y Gerstein, MB 2007. Análisis estadístico de la distribución genómica y correlación de elementos reguladores en las regiones ENCODE. *Genoma Res.* (este número) doi: 10.1101/gr.5573107. Zheng, D. y Gerstein, MB 2007. El límite ambiguo entre Genes y pseudogenes: Los muertos se levantan, ¿o no? *Tendencias Genet.* **23:** 219–224.

Zheng, D., Zhang, Z., Harrison, PM, Karro, J., Carriero, N. y Gerstein, M. 2005. Anotación pseudogénica integrada para el cromosoma 22 humano: Evidencia de transcripción. *J. Mol. Biol.* **349:**27–45. Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, SW, Lu, Y., Denoeud, F., Antonarakis, SE, Snyder, M., et al. 2007. Pseudogenes en las regiones ENCODE: anotación de consenso, análisis de transcripción y evolución. *Genoma Res.* (este número) doi: 10.1101/gr.5586307.

¿Qué es un gen, post-ENCODE? Historia y definición actualizada

Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, et al.

Genoma Res.2007 17: 669-681
Accede a la versión más reciente en doi:10.1101/gr.6339607