Arbor

Secuenciación de genomas

Javier María Rodríguez Martínez

Arbor CLXXVII, 698 (Febrero 2004), 285-310 pp.

Desde que en 1995 se determinó la secuencia del genoma del primer organismo autosuficiente, la bacteria Haemophilus influenzae, estamos asistiendo a una explosión en el número de genomas secuenciados. A finales del 2003 este número era de 150 y probablemente se doblara durante este año. También en 2003 se hizo pública la Secuencia de Referencia del genoma humano, un genoma de particular importancia para las ciencias biomédicas, y cuyo proyecto internacional de secuenciación ha sido el principal motor para el desarrollo de las tecnologías necesarias para este crecimiento. La posibilidad de analizar y comparar entre si toda la información genética de diversos organismos esta produciendo una rápida transformación de las ciencias biomédicas. En este articulo describiremos los métodos de secuenciación de genomas complejos que han hecho posible esta revolución y que suponen la base del conjunto de técnicas y conocimientos que conocemos como genómica.

Introducción

Podemos definir la genómica como la subdisciplina de la genética interesada en la descripción y análisis molecular de genomas completos. Habitualmente la genómica se suele subdividir en dos grandes áreas: La *genómica estructural*, que se ocupa de la caracterización de la naturaleza física de los genomas, y la *genómica funcional*, cuyo objetivo último es ubicar todos los elementos integrantes de un genoma dentro de una estructura funcional, tanto en el sentido más tradicional de determinar la función de cada una de los elementos componentes de un genoma (las

proteínas codificadas, los elementos reguladores, estructurales, etc) como en el sentido más general de determinar el papel que cada uno de estos elementos desempeña en el funcionamiento global del organismo. La mayor parte de los proyectos de genómica se encuentran aún en la fase estructural, pero en el caso de algunos organismos modelo como la mosca del vinagre (*D. melanogaster*) o el nematodo (*C. elegans*), la fase funcional ya ha comenzado.

En este articulo describiremos las técnicas de lo que hemos denominado genómica estructural, esto es, el conjunto de métodos y herramientas diseñadas para la determinación de la secuencia de genomas, y nos centraremos fundamentalmente en las empleadas para la secuenciación de genomas complejos, como los de los organismos eucarióticos.

Material genético

Cada organismo, sea este un virus, una bacteria, un animal o una planta, posee un genoma que contiene la información biológica necesaria para construir y mantener cada una de las instancias de ese organismo. La mayor parte de los genomas presentes en la naturaleza están constituidos por ácido desoxirribonucleico (DNA) aunque ciertos virus poseen ácido ribonucleico (RNA) como material genético. Tanto el DNA como el RNA son moléculas poliméricas construidas por cadenas de subunidades denominadas nucleótidos, desoxirribonucleótidos en el caso del DNA (de ahí la D), y ribonucleótidos en el caso del RNA. El DNA está compuesto por una mezcla de cuatro de estos nucleótidos: la adenina, que se representa con una A, la guanina (G), la citosina (C) y la timidina (T). Una molécula de DNA esta formada por dos cadenas de estos nucleótidos polimerizados, que se denominan bases, formado una estructura que se describe a menudo como una doble hélice. Las dos cadenas o hebras del DNA están estabilizadas entre si por puentes de hidrógeno, que ocurren entre las bases de las dos cadenas. Decimos que las bases están apareadas unas con otras. Este apareamiento tiene lugar de una forma muy precisa: la A de una cadena se aparea con la T de la otra cadena y la C con la G. La información biológica presente en el DNA se encuentra codificada en el orden preciso de esos nucleótidos dentro de la molécula de DNA, lo que denominamos secuencia de nucleótidos. El objetivo primario de la genómica estructural es precísamente determinar la secuencia de nucleótidos específica de cada genoma.

El Genoma

El humano es un buen ejemplo de genoma eucariótico complejo. Consiste en dos partes diferenciadas, el genoma mitocondrial y el genoma nuclear. La mitocondrias en las células animales y los cloroplastos en las células de plantas son los únicos orgánulos subcelulares que poseen su propio «genoma». El genoma mitocondrial humano es una pequeña molécula de DNA circular de 16.569 nucleótidos. En una célula normal puede haber unos 200 de estos orgánulos, cada uno con su propia copia de su genoma. Sin embargo, la mayor cantidad de información genética del ser humano se encuentra en el genoma nuclear compuesto por aproximadamente 3.200 millones de nucleótidos. El genoma nuclear, que es lo que normalmente se denomina genoma humano, esta dividido en 24 moléculas lineales cada una de ellas contenidas en un cromosoma diferente. La más pequeña de estas moléculas tiene unos 50 millones de nucleótidos mientras que la mayor tiene aproximadamente 250 millones de nucleótidos.

En un humano adulto, cada una de las aproximadamente 10^{13} células que lo componen contiene su propia copia del genoma, con la excepción de algunas células muy especializadas como los glóbulos rojos que en su estado final, completamente diferenciado, carecen de núcleo. La inmensa mayoría de las células contienen dos copias de cada uno de los cromosomas, solamente las células germinales (espermatozoides y óvulos) poseen un solo juego de cromosomas.

La secuencia del genoma mitocondrial humano fue determinada en 1981 y sin embargo hasta el 2003 no ha sido posible hacer lo propio con la secuencia del enormemente complejo genoma nuclear.

Secuenciación del DNA

La técnica empleada en la actualidad para la secuenciación de DNA es una modificación de la desarrollada en los años 70 por Frederick Sanger y colaboradores, conocida como el método de los terminadores de cadena. Esta técnica (Figura 1) se basa en el empleo de una enzima, DNA polimerasa, cuya actividad principal es la de extender una cadena de DNA polimerizando nucleótidos en uno de sus extremos. Para su funcionamiento esta enzima necesita esencialmente tres reactivos: un DNA que le sirve de molde, otro DNA que le sirve de iniciador de la reacción (en uno de cuyos extremos adicionara los nucleótidos) y los 4 nucleótidos componentes del DNA. La clave de esta técnica consiste en adicionar, junto a los reactivos

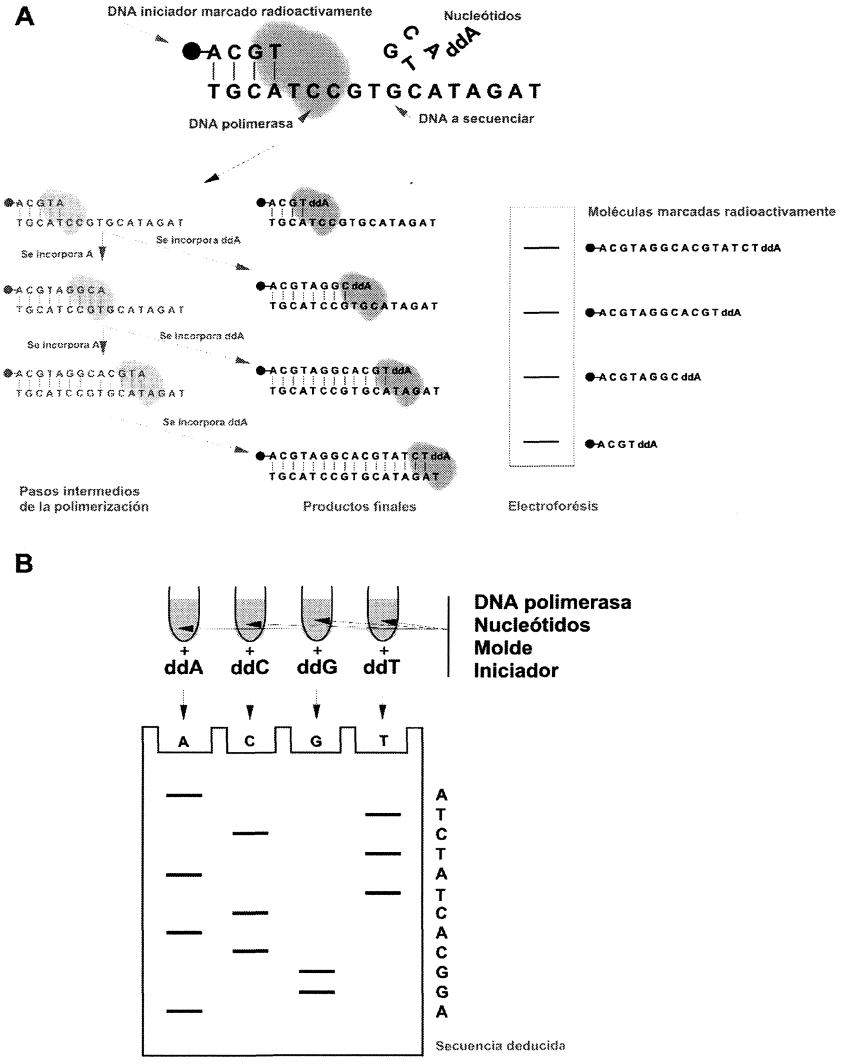

FIGURA 1. Secuenciación del DNA mediante el método de Sanger o de los terminadores de cadena. El panel A muestra un esquema de la reacción correspondiente a la determinación de la posición en la secuencia del DNA de una base, en este caso la A. La incorporación de ddA en lugar de A detiene el proceso de polimerización. Esta incorporación ocurre de forma aleatoria durante la polimerización de las moléculas de DNA de tal forma que una fracción de las moléculas elongadas se habrá detenido en cada posición en que A debiera incorporarse a la molécula. Las moléculas marcadas radioactivamente se detectan después de haberlas separado por su tamaño mediante electroforésis en geles de poliacrilamida. El tamaño de las moléculas detectadas nos indica en que posición de la secuencia se encuentra el núcleotido A. En el panel B se muestra un experimento completo de secuenciación del DNA. Se realizan reacciones como las descritas anteriormente para cada uno de los nucleótidos y la secuencia completa del DNA se deduce de la posición en que aparecen las moléculas marcadas radioactivamente.

mencionados anteriormente, una pequeña cantidad de nucleótidos modificados que se incorporan en la cadena que se esta elongando haciendo imposible que la polimerización en esta molécula continúe, es decir, que actúan como terminadores de la cadena. Por ejemplo, si en una reacción añadimos un porcentaje del nucleótido A modificado (Figura 1A), que denominaremos ddA, en cada una de las posiciones en las que se debe incorporar una A, una fracción de las moléculas que se están sintetizando incorporaran en su lugar ddA y la polimerización se detendrá en este nucleótido. En el resto de las moléculas, en las que se ha incorporado correctamente una A, la polimerización continuará hasta la siguiente A de la secuencia, momento en que se repetirá la situación anterior, una fracción de las moléculas incorporará ddA deteniéndose la reacción de polimerización en estas moléculas, y el resto continuará con el proceso de polimerización. Esta situación se repetirá en cada posición donde se deba incorporar una A en la secuencia. La incorporación de ddA en lugar de A ocurre de forma aleatoria, por lo que una fracción de las moléculas que se están sintetizando se detendrán en cada posición donde existe una A en la secuencia. Al final de la reacción obtendremos una mezcla de moléculas de diferentes tamaños que han resultado de las paradas de la polimerización en todas las posiciones donde existe una A en la molécula. Si separamos estas moléculas según su tamaño (empleando técnicas de electroforésis) podemos deducir, por su tamaño, en que posiciones se ha parado la polimerización de una parte de las moléculas y, por lo tanto, en que posiciones existe una A en la secuencia de ese DNA. En el método original de Sanger, la detección de las moléculas de DNA en la reacción de secuenciación se realiza utilizando un DNA iniciador marcado radioactívamente.

Para obtener la secuencia completa de una molécula de DNA (Figura 1B) lo que hacemos es correr en paralelo reacciones como la descrita anteriormente para los cuatro componentes del DNA. En ellas se añaden los mismos reactivos y un nucleótido modificado diferente (ddA, ddG, ddC o ddT), según cual sea el tipo de bases que queremos determinar en esa reacción. Las moléculas de DNA sintetizadas en cada una de las reacciones se separan en paralelo mediante electroforésis y la secuencia de la molécula de DNA se deduce observando en que reacción se ha parado la elongación correspondiente a esa posición.

Con esta técnica se pueden leer alrededor de 300 - 500 nucleótidos en cada experimento. Para secuenciar una molécula de mayor tamaño, tendremos que utilizar iniciadores diferentes que comiencen la reacción de polimerización en posiciones separadas unos 300 nucleótidos entre si.

Modificaciones posteriores de esta técnica eliminaron la necesidad de emplear iniciadores marcados, usando en su lugar uno de los nucléotidos

marcados radioactivamente de tal forma que las moléculas se marcan a medida que se van elongando.

Bajo la presión del Proyecto Genoma Humano por desarrollar nuevas tecnologías que permitieran la determinación de la secuencia del DNA con una mayor rapidez, esta técnica sufrió una serie de modificaciones dando lugar a un método mas sólido y sobre todo, susceptible de un gran nivel de automatización. Estas modificaciones afectaron fundamentalmente a:

(i) Mejoras en los reactivos bioquímicos necesarios para las reacciones de secuenciación, como polimerasas termoestables, terminadores marcados con colorantes fluorescentes y mejoras posteriores de la estabilidad de estos colorantes. La aparición, en 1986, de terminadores marcados con colorantes fluorescentes permite la realización de una reacción de secuenciación en un solo tubo, en lugar de los cuatro que eran necesarios en la técnica original de Sanger. Esto es posible porque cada uno de los cuatro terminadores de cadena esta marcado con un colorante diferente, permitiéndonos diferenciar, por el tipo de fluorescencia, que terminador se ha incorporado a cada molécula.

(ii) Desarrollo de secuenciadores que permiten la lectura automática del resultado de la reacción, al emplear terminadores fluorescentes.

(iii) Desarrollo de secuenciadores basados en electroforésis capilar para la separación de las reacciones. Estos aparatos permiten el procesamiento simultaneo de un número mucho mayor de reacciones de secuencia que los anteriores, basados en electroforesis en geles de poliacrilamida.

(iv) Desarrollo de sistemas robóticos para la automatización de la mayor parte de las tareas a realizar, como el aislamiento de clones, crecimiento y purificación del DNA, y la preparación de las reacciones de secuencia e incluso la colocación de estas reacciones en los secuenciadores automáticos.

En conjunto, todas estas mejoras han conseguido una mayor precisión en las lecturas y una automatización prácticamente completa del proceso. Por otro lado la construcción de grandes centros dedicados casi exclusivamente a la secuenciación ha permitido una disminución de los costes debido a la centralización y al aumento de la escala. En conjunto toda esta serie de mejoras han llevado a una reducción de mas de 100 veces en el coste por base secuenciada en la última década.

Un ejemplo de los elevados niveles de automatización obtenidos en la secuenciación de DNA lo proporciona la empresa Celera Genomics. Durante el período de secuenciación del genoma de la mosca del vinagre y

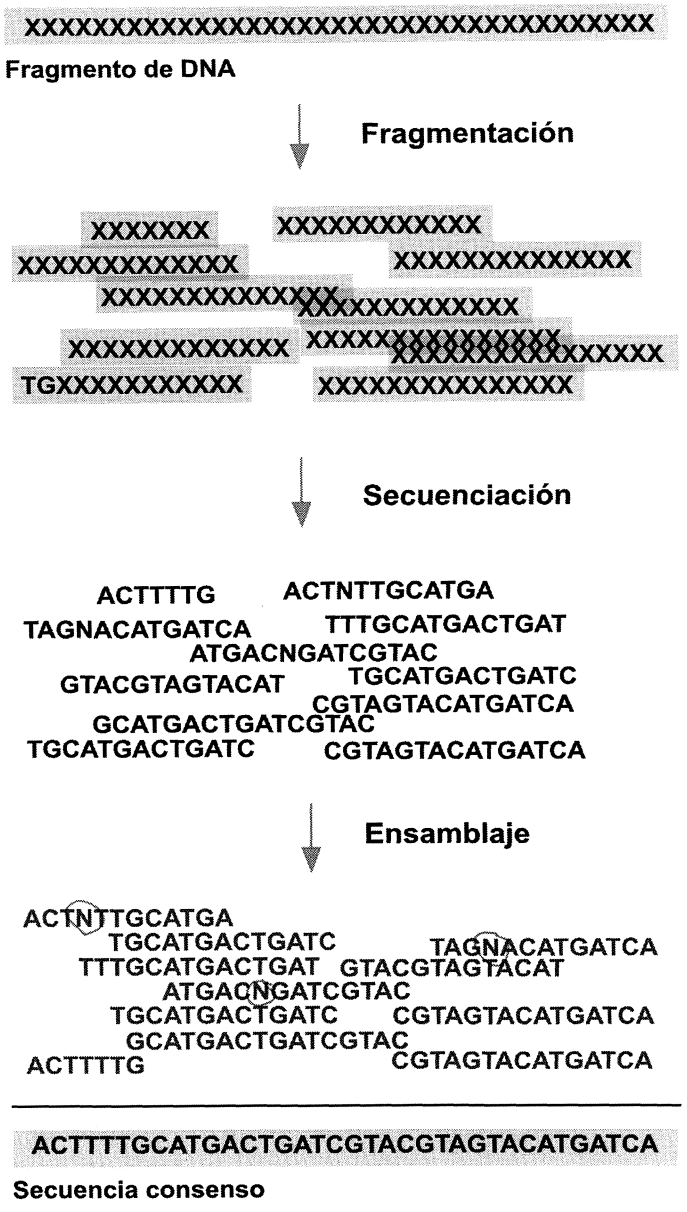

FIGURA 2. Esquema del proceso de secuenciación al azar. El fragmento de DNA de secuencia desconocida se rompe de forma aleatoria en fragmentos más pequeños que se procesan hasta obtener una cantidad de secuencia equivalente a varias veces la de la molécula original. Las secuencias de los fragmentos se ensamblan, reconstruyéndose así la secuencia de la molécula original. Gracias a esta redundancia podemos corregir con facilidad los errores que se hayan producido en las secuencias individuales.

del genoma humano, esta empresa producía diariamente 175000 lecturas, con un total de unos 95 millones de bases identificadas. Su nivel de automatización permitía que el tiempo real dedicado por cada operario a un secuenciador fuera de unos 15 minutos y el grado de integración obtenido entre los distintos departamentos (preparación de clones, obtención de DNA, preparación de reacciones y determinación de la secuencia) permitió mantener este nivel de producción durante varios años sin un solo día de interrupción.

Estrategias para la secuenciación de genomas complejos

Aunque a lo largo de casi 30 años las técnicas de secuenciación de ácidos nucleicos han sufrido importantes modificaciones, la limitación fundamental sigue siendo la cantidad de secuencia que es posible determinar en una sola reacción, lo que se denomina una lectura. Actualmente es de unas 1000 bases de las que 800 como máximo son de alta calidad. Para obtener la secuencia de una molécula de tamaño mayor que estos 800 nucleótidos es necesario ir empleando iniciadores separados entre si. Por tanto solo podemos realizar una reacción de secuenciación una vez que hemos llevado a cabo la anterior, determinado la secuencia y elegido el iniciador adecuado. Como vemos este proceso es extremadamente lento y difícilmente automatizable.

Otro problema importante es que estas secuencias poseen errores en un porcentaje de aproximadamente el 0.1 %. Esto es, 1 de cada 1000 bases puede ser errónea. Para soslayar estas limitaciones se han desarrollado una serie de estrategias para la secuenciación de grandes moléculas de DNA todas ellas basadas en la técnica de secuenciación al azar (*shotgun sequencing*), descrita por Sanger. En este método (Figura 2) el DNA a secuenciar se rompe de forma aleatoria en fragmentos más pequeños que se procesan hasta obtener una cantidad de secuencia equivalente a varias veces la de la molécula original, lo que denominamos redundancia. Gracias a esta redundancia podemos (1) ensamblar las secuencias de estos fragmentos para deducir la secuencia de la molécula original y (2) corregir con facilidad los errores que se hayan producido en las secuencias de los fragmentos ya que para cada posición de la molécula original tenemos varias secuencias redundantes.

Si asumimos que la fragmentación ha sido realmente aleatoria, la fracción de genoma que permanece sin secuenciar en ambas cadenas se puede calcular como

$$p_0 = e^{-n\omega/L}$$

donde n es el número de los fragmentos secuenciados, ω es el tamaño medio de estos fragmentos y L es el tamaño de la molécula original expresado en miles de bases (kb). Una redundancia de 9 ($nw/L = 9$) producirá aproximadamente el 99,99 % (po = 0.01) de la secuencia original suponiendo una distribución realmente aleatoria. Sin embargo en la realidad existen muchos factores, que esta versión idealizada no contempla, que hacen necesarias redundancias incluso mayores. Como se observa, el mayor inconveniente de esta técnica es que para obtener la secuencia de nucleótidos de una molécula de 5000 bases es necesario secuenciar un número total de 45000 bases.

Esta estrategia se puede aplicar en principio a cualquier molécula, sin importar su tamaño, siempre que no contenga secuencias repetidas y que podamos fragmentarla al azar. Si esto es así, el ensamblaje de las secuencias de los fragmentos requiere programas informáticos relativamente sencillos. Los problemas prácticos de esta técnica provienen de las secuencias repetidas presentes en los genomas y de las desviaciones del azar que se producen durante la preparación de los fragmentos del genoma. En el caso de las repeticiones, un número pequeño de ellas tampoco plantea una gran dificultad. Por ejemplo se han ensamblado sin problemas genomas bacterianos típicos que contienen un 1.5 % de secuencias repetidas, o la porción eucromática del genoma de la mosca del vinagre, que contiene un 3% de secuencia repetidas. Sin embargo, el genoma humano, por ejemplo, contiene mas de un 50% de secuencia repetidas que incluyen grandes fragmentos, resultantes de duplicaciones, con una similitud de secuencia del 98 al 99.9 %. Otros genomas como los de las plantas contienen una cantidad muy superior de secuencias repetidas. Estas características complican considerablemente el ensamblaje de la secuencia completa de estos genomas ya que con una similitud del 99.9%, y teniendo en cuenta los posibles errores de las lecturas, es prácticamente imposible para el programa empleado en el ensamblaje discriminar la posición correcta de una lectura que es 99.9% idéntica a dos secuencias repetidas.

Se han empleado dos estrategias para la secuenciación de genomas con repeticiones: la secuenciación al azar jerárquica (*hierarchical shotgun sequencing*) y la secuenciación al azar de todo el genoma (*whole-genome shotgun sequencing*). Un tercera estrategia, una especie de híbrido de las dos anteriores y que incorpora las mejores características de ambas, es la que en la actualidad parecen preferir los grandes proyectos de secuenciación, como los del genoma de la rata y del ratón. En el caso de los proyectos de secuenciación de genomas extremadamente ricos en repeticiones, como el del maiz, los esfuerzos se han centrado en desarrollar

técnicas que permitan discriminar entre las regiones con DNA puramente repetitivo y las regiones que contienen DNA no repetitivo, rico en genes, para obtener la secuencia únicamente de este último

Secuenciación al azar jerárquica

La característica fundamental de esta estrategia es la obtención, previa a la secuenciación, de un mapa del genoma mediante grandes fragmentos de DNA de unos 100 a 200 kb. En este mapa, cada posición en el genoma esta representada en varios fragmentos, es decir la colección de fragmentos posee una elevada redundancia. Antes de comenzar la secuenciación se eligen, entre los componentes del mapa, una serie de fragmentos que solapen entre si y que abarquen todo el genoma. Estos se secuencian por el método secuenciación al azar y las secuencias individuales de los fragmentos se ensamblan siguiendo tanto el mapa físico previamente construido como las regiones de solapamiento detectadas, generándose de esta forma la secuencia completa del genoma. Un esquema de este método se puede ver en la figura 3. Esencialmente, la idea es fragmentar un problema grande en pequeños problemas que podemos resolver fácilmente, sumar las soluciones de estos pequeños problemas y obtener la solución a nuestro gran problema original. Conceptualmente, este método se puede dividir en una serie de pasos:

1. *Construcción del mapa físico*

La construcción del mapa físico (Figura 3) comienza con el aislamiento del DNA genómico. Este DNA se rompe en fragmentos de unos 150 kb mediante métodos físicos o empleando enzimas que lo cortan. Para poder aislar, amplificar y almacenar estos fragmentos de DNA es necesario introducirlos en vectores adecuados, proceso que se denomina clonaje. Para ello se han empleado dos sistemas fundamentales, los vectores desarrollados a partir de levaduras, YAC (siglas en ingles para cromosomas artificiales de levadura) y los vectores desarrollados a partir de bacterias, BAC (siglas del ingles para cromosomas artificiales de bacterias) o PAC (siglas del ingles para cromosomas artificiales derivados de P1) muy parecidos a los BAC. Los YAC son capaces de aceptar fragmentos de hasta 1 Mb (un millón de bases) y se emplearon para realizar los mapas físicos de primera generación de los genomas de ratón y humano. Sin embargo por diversas razones técnicas y por la inestabilidad del DNA insertado en estos vectores, los YAC no son buenos puntos de partida para los si-

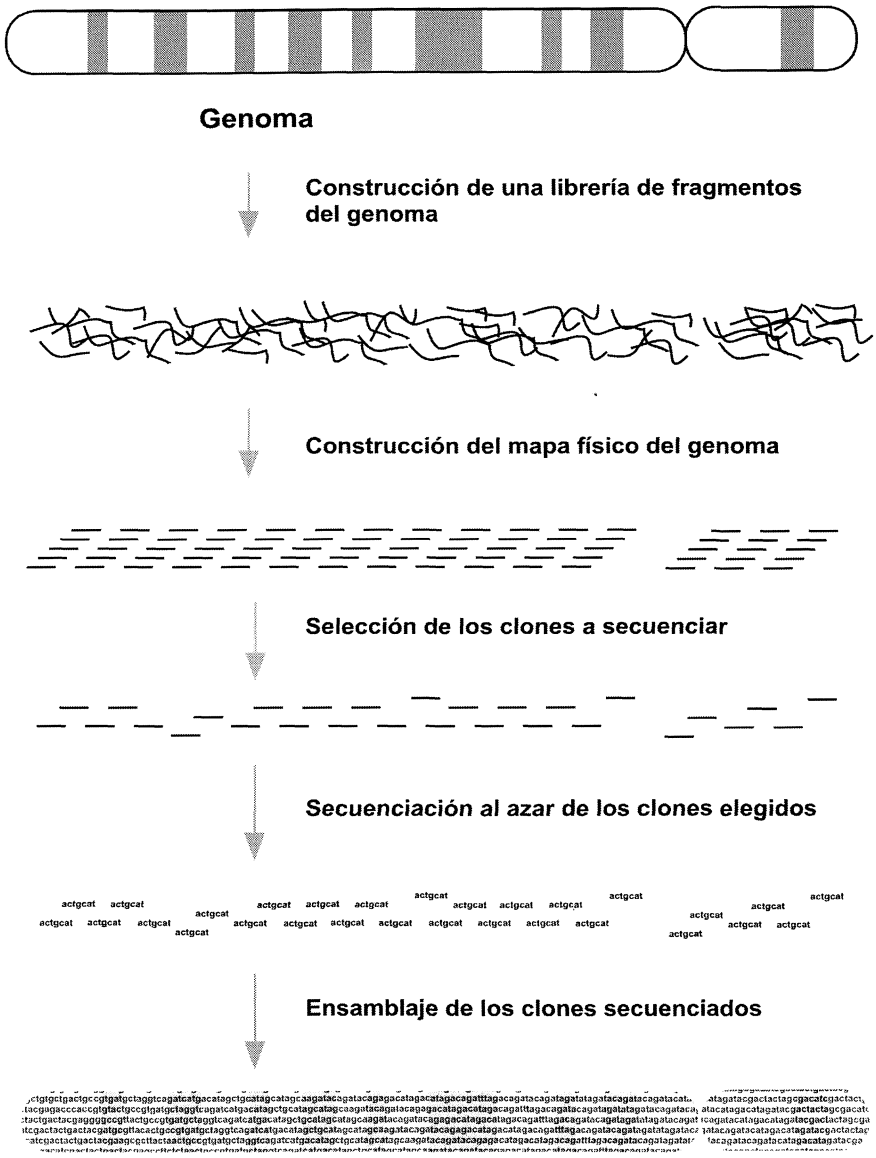

FIGURA 3. Esquema del método de secuenciación al azar jerárquica. El DNA del genoma se rompe en fragmentos de gran tamaño que son ordenados según su posición en el genoma. De entre estos clones se selecciona una colección que abarca todo el genoma. A continuación se obtiene la secuencia de estos fragmentos individuales empleando la técnica de secuenciación al azar. Finalmente la secuencia del genoma se reconstruye ensamblando la secuencia de estos fragmentos.

guientes pasos del proceso. En general los vectores basados en BAC son los más empleados para la generación de mapas físicos. En ellos es posible introducir un fragmento de DNA de unos 100-200 kb. Para la secuenciación del genoma humano, el Consorcio Internacional para la Secuenciación del Genoma Humano empleo 8 librerías de clones basados en vectores BAC y PAC con un tamaño medio de unos 150 kb.

El siguiente paso es ordenar los clones de estas librerías según su posición en el genoma (Figura 4). Para ello se emplean diferentes técnicas que, en general, implican la identificación de ciertos marcadores característicos (pequeñas secuencias únicas (STS), sitios de corte de enzimas de restricción etc...) en cada uno de los fragmentos clonados. Mediante la comparación de la presencia de dichos marcadores en los diferentes clones, estos son ordenados de forma inequívoca. En general se escoge un número de clones elevado para que la misma zona del genoma este representada en varios de ellos, es decir tener una alta redundancia de la secuencia del genoma en la librería de BAC. En el caso del Proyecto Genoma Humano, el mapa generado mediante los clones de las librerías de BAC y PAC tenia una redundancia de 65 veces el genoma.

2. *Selección de los clones*

Una vez obtenido un mapa del genoma mediante la ordenación de los clones de las librerías, se escoge el número mínimo de clones en los que este contenido todo el genoma, minimizando las zonas de solapamiento ente ellos. En este momento es crítico elegir clones que no hayan sufrido anomalías durante el proceso de construcción de la librería, como pueden ser la pérdida de parte del fragmento de DNA clonado (deleciones) o la presencia de dos fragmentos de diferentes zonas del genoma clonados en el mismo vector (clones quiméricos), puesto que son los clones de los que vamos a obtener la secuencia final. Para reducir estos problemas al mínimo, los marcadores de los BAC candidatos se comparan con los de los otros clones que abarcan la misma zona, eligiéndose aquellos en los que los marcadores concuerdan.

3. *Construcción de las librerías al azar de subclones*

El siguiente paso es la secuenciación de los clones de BAC elegidos como representantes de cada una de las regiones de aproximadamente

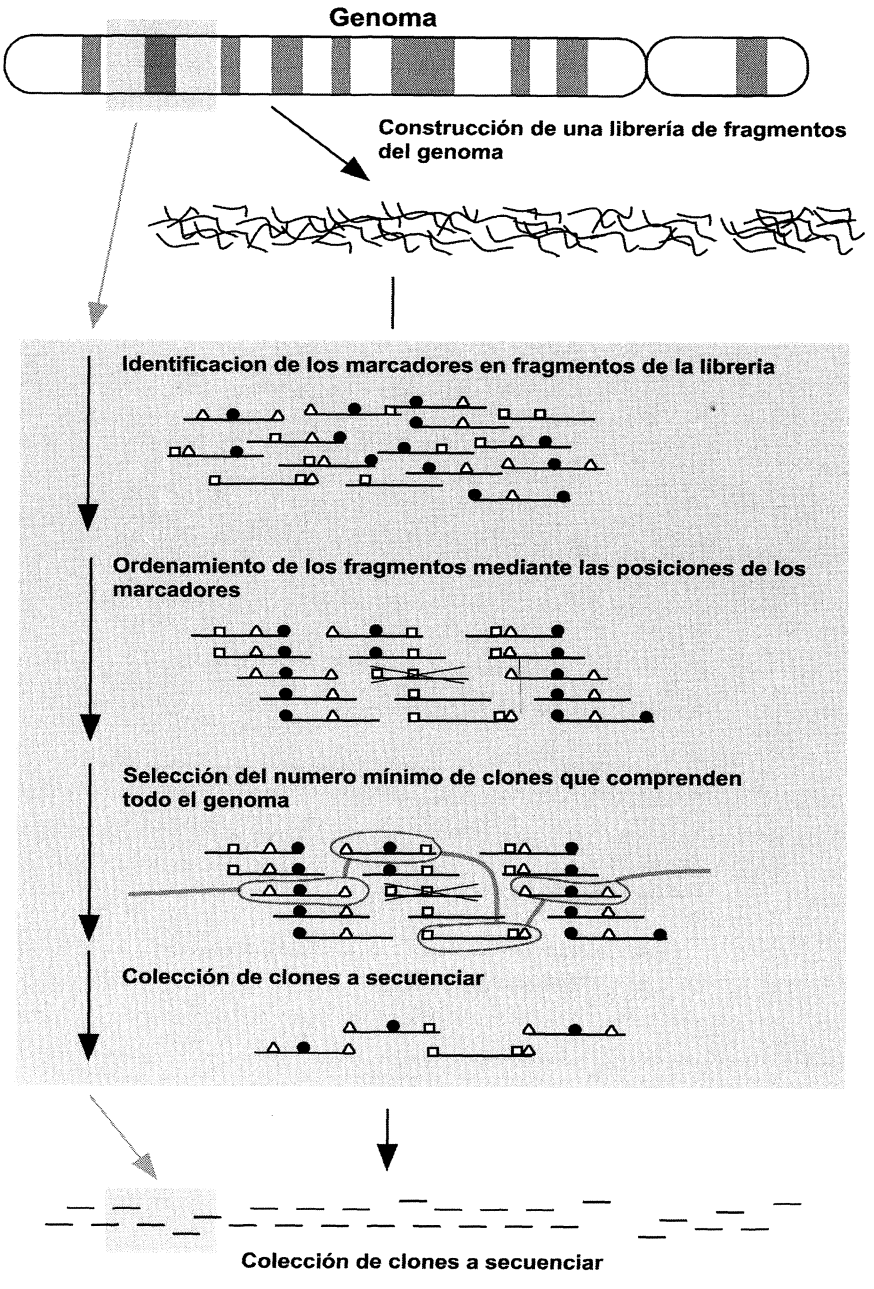

FIGURA 4. Secuenciación al azar jerárquica: Construcción del mapa de clones. En la figura se muestran gráficamente los pasos seguidos para la construcción de un mapa de clones de un genoma

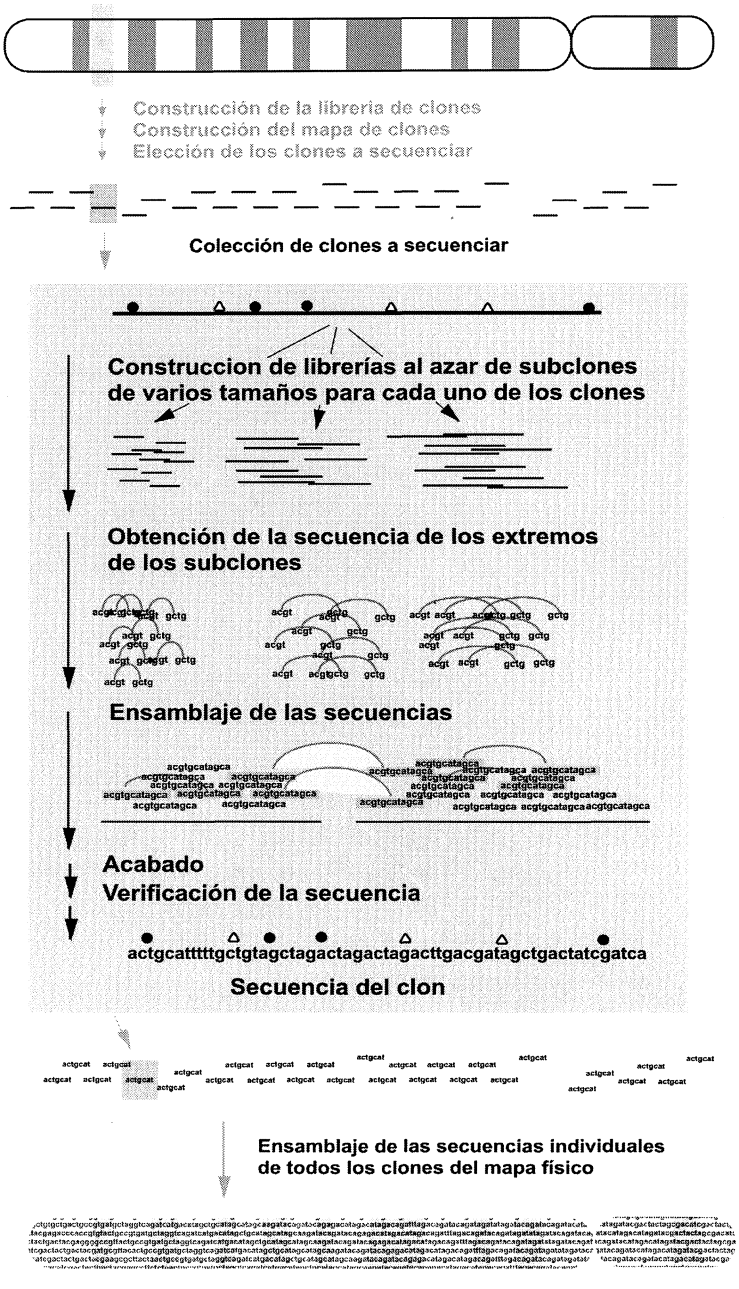

FIGURA 5. Secuenciación al azar jerárquica: Secuenciación de los clones del mapa. En la figura se muestran gráficamente los pasos seguidos para la secuenciación y el ensamblaje de la secuencia de los clones del mapa físico del genoma

150 kb en que hemos dividido el genoma (Figura 5). Este proceso comienza con la purificación del DNA de cada uno de los clones seleccionados y su posterior fragmentacion al azar, en general mediante métodos físicos como la sonicación o el paso forzado por pequeños orificios a gran presión. Los fragmentos resultantes se separan por tamaños y los que se encuentran en un rango adecuado, por ejemplo de 2 a 5 kb, son posteriormente clonados en vectores derivados del bacteriófago M13 o de plásmidos. La ventaja de estos últimos vectores es que el fragmento de doble cadena de DNA clonado puede ser secuenciado por sus dos extremos (al coste de una única preparación de DNA) y que las dos lecturas derivadas de cada uno de ellos (lo que se conoce como pareja de lecturas) puede ser usada para facilitar y/o verificar el proceso de ensamblaje. Esto es así porque conocemos la distancia a la que deben encontrarse las dos secuencias de cada pareja de lecturas en la secuencia final, que además deberá coincidir con el tamaño del fragmento clonado en el plásmido de donde se han obtenido. Por otro lado, los vectores derivados del bacteriófago M13 tienen la ventaja de que el DNA es más fácil de preparar y de que el molde resultante, de cadena sencilla produce unas secuencias de mayor calidad. Ambos tipos de vectores, plásmidos y bacteriófagos, provocan cierta selección en las secuencias clonadas, siendo más fácil clonar cierto tipo de secuencias en unos que en otros. Estas desviaciones del azar deben ser minimizadas si quiere obtenerse una representación realmente aleatoria de la secuencia original del BAC. Para evitarlo, en ciertos centros, se generan simultáneamente ambos tipos de librerías minimizándose este tipo de problemas pero incrementándose la complejidad del proceso de secuenciación, al tener que preparar dos tipos de librerías y de purificar dos tipos de clones.

4. *Secuenciación al azar*

El grueso de la secuenciación se realiza sobre las librerías de subclones anteriormente citadas. Para ello se seleccionan aleatóriamente una serie de subclones, se prepara su DNA, y se determina la secuencia del extremo o de los extremos (dependiendo de si los subclones son plásmidos o M13 respectivamente) del fragmento clonado. Este proceso de secuenciación de subclones al azar continúa hasta generar una cantidad suficiente de secuencia redundante (con relación al inserto presente en el BAC original). Es ese momento se ensamblan las secuencias de los subclones mediante programas de ordenador gracias a los solapamientos detectados. Normalmente el resultado del ensamblaje es una serie ordena-

da de segmentos del fragmento original, que se denominan *contigs*, cada uno formado por una colección de lecturas solapantes. A partir de las bases presentes en cada posición en las lecturas solapantes es posible deducir una secuencia, que se denomina secuencia consenso.

Para producir una secuencia con una precisión superior al 99.99%, que es el estándar del Proyecto Genoma Humano, es necesario generar lecturas que supongan mas de 10 veces la cantidad de secuencia que queremos obtener (o sea una redundancia de mas de 10). Por ejemplo en el caso de un BAC de 150 Kpb son necesarias 3000 lecturas útiles (esto es descartando aquellas que no produjeron datos válidos, las derivadas de contaminaciones como secuencias del BAC, del vector empleado en la generación de los subclones y otras secuencia contaminantes) de unas 500 bases de calidad para obtener una redundancia de 10 veces. Cuando se alcanza este nivel de redundancia finaliza la fase de secuenciación al azar.

5. *Fase de secuenciación dirigida*

El ensamblaje de las lecturas con una redundancia de 10 genera una serie de *contigs* que, en conjunto, reflejan prácticamente la totalidad del clon inicial. Los problemas que quedan son generalmente discontinuidades entre los contigs, áreas donde la calidad de la secuencia es demasiado baja para el estándar elegido, bases individuales que permanecen ambiguas y zonas donde el ensamblaje de los contigs ha sido erróneo. En general, estos problemas se resuelven mediante la secuenciación adicional de subclones concretos así como con la secuenciación directa del DNA del BAC mediante oligonucleótidos específicos. A menudo es necesario el empleo de químicas de secuenciación diferentes a las empleadas en la secuenciación al azar, diseñadas para evitar cierto tipo de problemas derivados de la composición del DNA. En contraste con la automatización de la fase de secuenciación al azar, esta fase de acabado es un proceso lento y complejo que requiere mucha mayor atención por parte del investigador.

6. *Verificación de la secuencia*

Una vez terminado el ensamblaje se analiza la secuencia generada para determinar las presencia y el orden correcto de los marcadores conocidos de ese clon (como STS, sitios de corte de enzimas de restricción o

genes previamente localizados en esa región). Este paso es crucial para poder detectar errores cometidos en cualquiera de los procesos de la determinación de la secuencia del clon.

7. *Ensamblaje de la secuencia del genoma*

Finalmente, y siguiendo el orden determinado durante la elaboración del mapa físico, las secuencias de los clones BAC se ensamblan para generar la secuencia completa del genoma.

Este método de secuenciación se ha empleado para la obtención de la secuencia completa de los genomas de la levadura *S. cerevisae*, el nematodo *C. elegans* y la planta *A. thaliana*. Sin embargo, su uso mas notable ha sido en la obtención de la secuencia completa del genoma humano realizado por el Proyecto Genoma Humano, que el 14 de Abril de 2003, 50 años después de que Watson y Crick determinaran la estructura del DNA, dio por finalizada la secuencia. Esta proyecto de secuenciación se ha realizado mediante un riguroso proceso de secuenciación jerárquica al azar. La Secuencia de Referencia posée un nivel de precisión elevado (menos de un error por cada 10 000 bases) y comprende alrededor del 99 % de la secuencia total del genoma, correspondiendo los únicos vacíos restantes a regiones de los centrómeros y telómeros que, con las técnicas actuales, se consideran imposibles de clonar y secuenciar con fiabilidad.

Secuenciación al azar de todo el genoma

La estrategia de secuenciación al azar de todo el genoma es más sencilla conceptualmente. En esta estrategia (Figura 6) el genoma completo se ensambla a partir de lecturas obtenidas al azar, eliminándose la necesidad de construir mediante clones de gran tamaño un mapa físico. Este procedimiento comienza con la purificación y rotura al azar del DNA del genoma que queremos secuenciar. Posteriormente se construyen librerías de fragmentos de al menos tres tamaños diferentes (por ejemplo de 2, 10 y 50 kb). El DNA de clones elegidos aleatóriamente de estas librerías se purifica y se obtiene la secuencia de los extremos de los fragmentos del genoma clonados. Este proceso de secuenciación continua hasta que se ha obtenido una elevada redundancia (mayor que en la estrategia de secuenciación jerárquica). En este método es clave, para poder evitar los problemas derivados de posibles secuencias repetidas en el genoma, obtener parejas de lecturas de los extremos de la mayor canti-

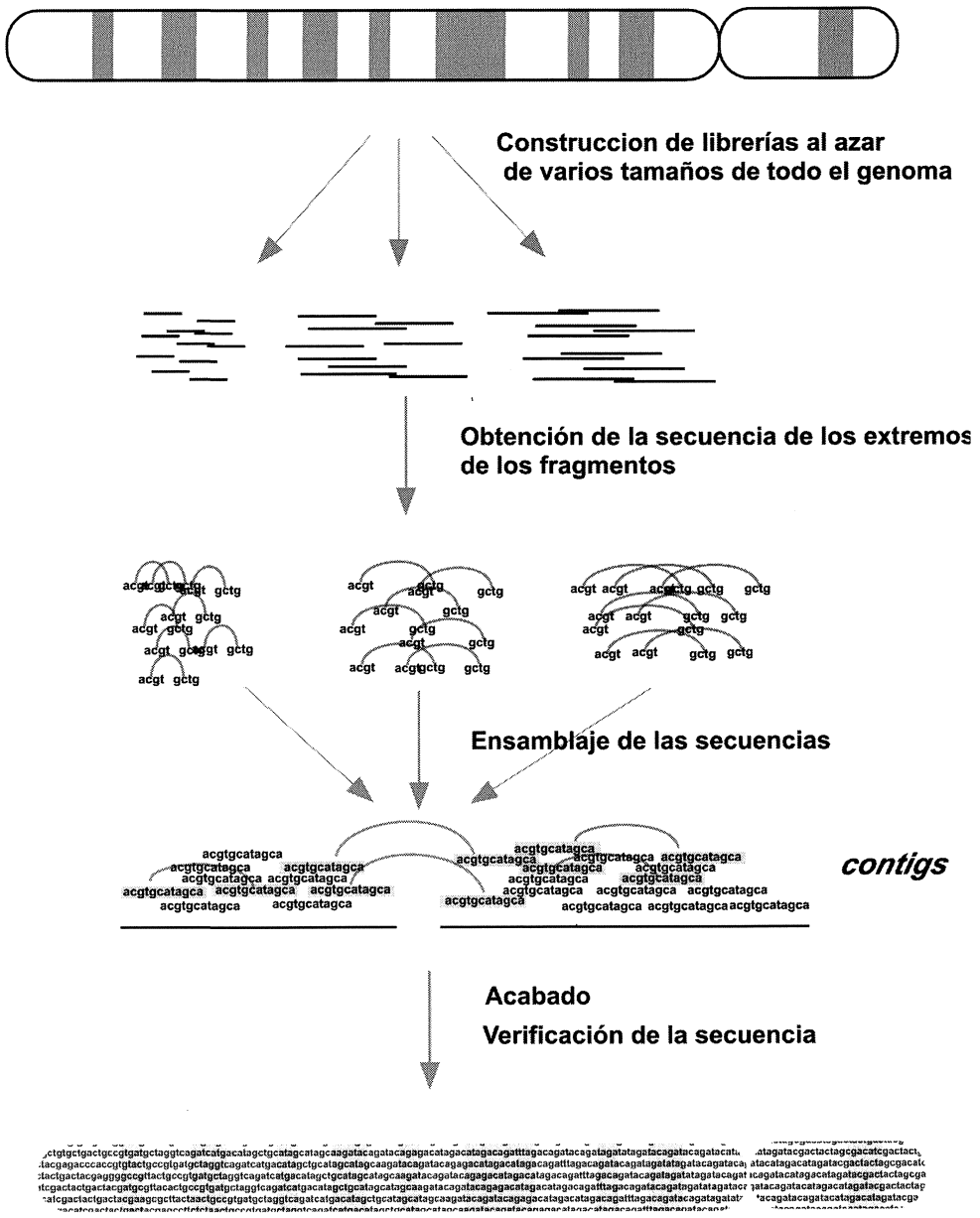

FIGURA 6. Secuenciación al azar de todo el genoma. En esta estrategia, el genoma completo se ensambla a partir de lecturas al azar obtenidas de todo el genoma, eliminándose la necesidad de construir un mapa físico a partir de clones de gran tamaño.

dad posible de fragmentos. Aquí, la distancia conocida entre las parejas de lecturas es esencial para el proceso de ensamblaje de las secuencias, ya que carecemos de cualquier otro tipo de información posicional. Una vez obtenida la redundancia requerida, se ensamblan las lecturas mediante potentes programas de ordenador, capaces de manejar un número muy elevado de lecturas (mas de 27 millones en el caso del ensamblaje del genoma humano).

Los proyectos de secuenciación realizados con esta técnica han puesto de manifiesto que, para el posterior éxito del ensamblaje de las secuencias, es esencial construir varias librerías de fragmentos de todo el genoma con tamaños muy diferentes, que cumplen diferentes misiones durante el proceso de ensamblaje por ordenador. Las librerías de pequeño tamaño (aproximadamente 2 kb) son sobre las que se realiza la mayor parte de la secuenciación. Las librerías de tamaño medio (aproximadamente 10 kb) suministran parejas de lecturas que son esenciales para la construcción de los *contigs* y para deducir la orientación y el orden de unos *contigs* con respecto a otros. Las librerías de gran tamaño (aproximadamente 50 kb) permiten obtener parejas de lecturas muy alejadas entre si que son necesarias para evitar los problemas derivados de bloques de secuencias repetidas, además de suministrar información a una escala mayor sobre la organización de los *contigs*.

Este método de secuenciación se aplico por primera vez a un organismo eucariota durante la secuenciación del genoma de *D. melanogaster*. En este proyecto se obtuvo prácticamente toda la secuencia de la parte eucromática del genoma. Sin embargo es importante señalar que la etapa final de refinamiento de la secuencia de este genoma se realizó mediante el uso de un mapa físico de clones de BAC.

La aplicación más destacada de esta estrategia ha sido la secuenciación del genoma humano, realizada por la empresa Celera Genomics, que puso de manifiesto tanto las ventajas como las debilidades de esta estrategia para la secuenciación de genomas eucarióticos. La ventaja fundamental es la velocidad con la que se obtiene una gran cantidad de secuencia, suficiente como para tener una idea bastante aproximada de la práctica totalidad del genoma, al no ser necesario el paso previo de construcción de un mapa de clones. Celera Genomics realizo 27.271.853 secuencias con un total de 14.8 10^9 bases leídas, aproximadamente 5.11 veces la secuencia del genoma humano en solo 9 meses. Sin embargo, para el posterior ensamblaje de estas lecturas y sobre todo para la localización de los *contigs* resultantes tuvo que recurrir a datos externos, fundamentalmente a los mapas físicos del genoma previamente realizados por el Proyecto Genoma Humano. Por otro lado Celera renunció al acabado de

la secuencia, probablemente porque para realizarlo hubiera sido necesario disponer, como en el caso de *D. melanogaster*, de un mapa físico del genoma realizado mediante BACs.

Esta estrategia se emplea de forma rutinaria para la secuenciación de genomas de organismos procariotas relativamente pequeños (0.5 a 6 Mb) y con pocas repeticiones. Se empleó por primera vez para la secuenciación del genoma de la bacteria *Haemophilus influenciae.*

Método híbrido

Este método pretende aprovechar lo mejor de las dos técnicas descritas anteriormente: la rapidez del método de secuenciación al azar para obtener una gran cantidad de secuencia y la capacidad del método jerarquizado para minimizar la influencia de las repeticiones, junto con la ventaja de poseer de un mapa físico del genoma realizado mediante BACs para el posicionamiento correcto de los contigs y el proceso de acabado de la secuencia (Figura 7).

En esta estrategia se comienza secuenciando el genoma mediante la modalidad de secuenciación al azar de todo el genoma. Simultáneamente se construye una librería de clones y se realiza un mapa físico con ellos. De esta forma, con la secuenciación inicial podemos tener una idea aproximada de la organización del genoma, de la abundancia y características de las repeticiones presentes, así como de las dificultades que estas repeticiones pueden plantear en el ensamblaje. Una vez obtenido el mapa físico se seleccionan los clones de BAC adecuados y se procede a su secuenciación al azar. El ensamblaje se realiza de forma independiente para cada uno de los clones en que hemos dividido el genoma, como en la estrategia de secuenciación al azar jerárquica. La diferencia en este método es que a las lecturas procedentes de cada clon se les unen las procedentes de la secuenciación al azar de todo el genoma que corresponden al fragmento que vamos a ensamblar. Para identificar estas últimas, todas las lecturas del proyecto de secuenciación al azar de todo el genoma se comparan con las lecturas procedentes del clon BAC y se adicionan aquellas que solapan. De esta forma se aumenta la redundancia en el ensamblaje de cada uno de los clones. El proceso de acabado se realiza de igual forma que en el método de secuenciación jerárquica.

Existe un consenso sobre la necesidad de una redundancia de 8 a 10 veces la secuencia del genoma completo si se pretende conseguir una secuencia final de alta calidad. Sin embargo, en el método híbrido esta por

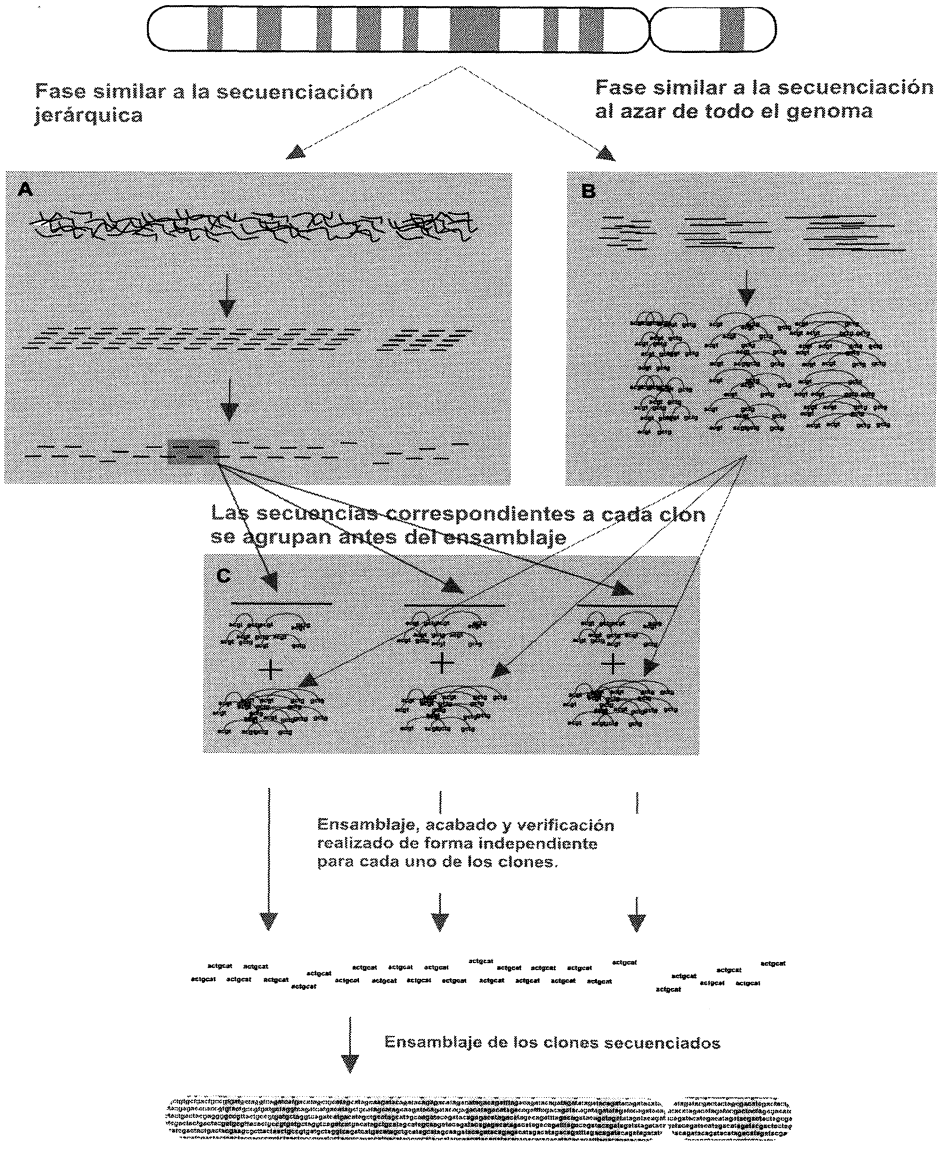

FIGURA 7. Método híbrido de secuenciación al azar de genomas. En este método se obtienen secuencias del genoma siguiendo las dos estrategias anteriores, la secuenciación al azar jerárquica (panel A) y la secuenciación al azar de todo el genoma (panel B). El ensamblaje (panel C) se realiza para cada uno de los clones en que se ha dividido el genoma en la parte jerárquica de este proceso juntando a las secuencias obtenidas de los clones, las secuencias correspondientes obtenidas del proceso al azar. El acabado de la secuencia de cada uno de los clones y la reconstrucción del genoma a partir de las secuencias individuales de los clones se realiza como en la secuenciación al azar jerárquica.

determinar la cantidad óptima de secuencia que es necesario conseguir mediante el proceso de secuenciación de todo el genoma frente a la que proviene de la secuenciación de los clones BAC. Los proyectos de secuenciación del genoma de la rata y del ratón, que se están llevando a cabo mediante esta técnica, es previsible que aporten datos sobre la relación optima de estas cifras.

Otras alternativas

A la hora de abordar la secuenciación de un genoma nuevo es muy importante evaluar exactamente el uso se le va a dar a la secuencia obtenida. Si el objetivo es obtener una secuencia de alta calidad, como en el caso de un organismo modelo, el proceso a seguir dependerá del tamaño, presencia y complejidad de las repeticiones del genoma. Para genomas pequeños con un número limitado de repeticiones la estrategia de secuenciación al azar directa es la más adecuada. Para genomas grandes con un elevado número de repeticiones la estrategia híbrida debe ser la elegida para garantizar la fiabilidad del proceso de acabado de la secuencia y minimizar el número de errores de ensamblaje debidos a la presencia de repeticiones.

Aunque siempre es deseable tener la secuencia completa de un organismo, el elevado coste de la secuenciación de un genoma complejo hace imposible disponer de las secuencias necesarias para el análisis comparativo de la secuencia de muchos genomas. La preparación de una secuencia de calidad media (lo que se denomina un borrador) de un genoma de un mamífero cuesta mas de 50 millones de euros. Por ello, para realizar estudios comparativos de genomas de muchos organismos es necesario restringirse a regiones concretas del genoma. Esta secuenciación dirigida se realiza mediante secuenciación al azar de clones de BAC que contienen las regiones de interés de los distintos organismos.

Otra técnica empleada cuando nos interesa realizar un estudio comparativo de los genomas de especies estrechamente relacionadas es lo que se denomina secuenciación de baja redundancia. Consiste en realizar un examen de los genomas de los diferentes organismos secuenciando un número aproximadamente 20 veces inferior de secuencias de las que seria necesario obtener para construir un borrador del genoma. Esta técnica presenta muchas limitaciones pero resulta extremadamente rápida y asequible para la caracterización y comparación de especies muy relacionadas entre si.

Más allá de la secuencia del genoma

Aunque en estos 30 años transcurridos desde la invención de los métodos de secuenciación la técnica se ha modificado considerablemente (disminuyendo, por ejemplo, el coste por base secuenciada en un factor de 100 veces en los últimos 10 años), estas modificaciones han sido básicamente mejoras de la técnica original, sin que se haya producido ningún cambio radical. Actualmente el esfuerzo se centra en el desarrollo de tecnologías que permiten disminuir drásticamente el volumen de las reacciones de secuenciación y los tiempos necesarios para realizar la separación electroforética de las moléculas. Se están explorando también métodos radicalmente distintos como la pirosecuenciación o la secuenciación basada en espectrometría de masas. Una verdadera revolución podría venir de la mano de la nanotecnología donde, mediante las tecnologías del campo de la biofísica de nanoporos, se están desarrollando proyectos dirigidos a obtener la secuencia completa de una única molécula de DNA. Estos métodos, aunque interesantes, están aún en fases tan preliminares que no permiten siquiera aventurar sus posibilidades prácticas en la secuenciación a gran escala.

Aunque la secuenciación del genoma es el objetivo fundamental de la genómica estructural, constituye el punto de partida necesario para comprender como funciona el genoma de un organismo. Habitualmente se habla del genoma como de el *«libro de la vida»*. Lo que nunca se dice es que tipo de libro es. Desde luego, no se trata de un manual de instrucciones fácilmente comprensible. Quizás un símil más adecuado fuera el de un libro de notas de un fabuloso ingeniero. Un cuaderno escrito en un lenguaje incomprensible, lleno de tachaduras, borrones, correcciones apresuradas y dibujos realizados mientras habla por teléfono. Dispersas entre ellas hay algunas anotaciones, sin ningún orden aparente, que describen con una precisión absoluta los componentes necesarios para construir y mantener en funcionamiento un organismo. Sin embargo en este cuaderno de notas no existe ninguna indicación comprensible sobre la forma en que todos estos componentes tienen que ensamblarse para que el organismo funcione.

El objetivo de la genómica funcional, para seguir con el símil anterior, es el de descifrar ese cuaderno de notas y construir, a partir de sus anotaciones, un «manual del usuario del genoma», comprensible para los humanos. Entre las tareas a las que se enfrenta la genómica funcional, tomando como ejemplo el caso del genoma humano, se encuentran:

i. **Identificar los componentes estructurales y funcionales del genoma**. Aunque la composición y características químicas

del DNA son bien conocidas, la estructura del genoma humano es extraordinariamente compleja. Únicamente un 1-2% de su secuencia codifica proteínas, y ni siquiera están identificadas con seguridad todas ellas. Aproximadamente una cantidad equivalente al doble del conjunto de secuencias codificantes se encuentra bajo presión evolutiva, lo que indica que son funcionalmente importantes, y sin embargo no conocemos prácticamente nada de sobre su función. Probablemente en ese otro 2% del genoma se encuentran los elementos que regulan la expresión de los aproximadamente 30.000 genes que codifican proteínas, junto con toda una serie de genes no codificantes y de secuencias determinantes de la estructura y funcionamiento de los cromosomas. Todavía se conoce menos sobre la posible función del aproximadamente 50% del genoma que consiste en secuencias repetidas, o del resto del genoma integrado por secuencias no codificantes y no repetidas.

ii. **Definir como interactúan los componentes del genoma a nivel genetico y protéico.** Los genes y sus productos, las proteínas, no actúan de forma aislada sino que forman parte de rutas, redes y sistemas que, en conjunto, dan lugar y mantienen en funcionamiento las células, los tejidos y los organismos. Para comprender como funciona un organismo es imprescindible entender el funcionamiento de estos sistemas y conocer sus propiedades e interacciones. Sin embargo, dichos sistemas como conjunto son mucho mas complejos que cualquier problema abordado antes por la biología molecular, la genética o la genómica.

iii. **Desarrollar un conocimiento detallado de la variación hereditaria en el genoma humano.** Los mayores avances en la genética humana se han producido sobre características hereditarias asociadas con modificaciones dependientes, en general, de un único gen. Sin embargo la mayor parte de los fenotipos, incluyendo enfermedades comunes o las respuestas a agentes farmacológicos, son mucho más complicados y dependen de una compleja interacción de factores genéticos (los genes y sus productos) y no genéticos (influencias ambientales). Para comprender como ocurre esta interrelación es necesario conocer la variación genética de la especie humana y desarrollar las herramientas analíticas necesarias para emplear este conocimiento en la determinación de las bases genéticas de las enfermedades.

iv. **Determinar los mecanismos causantes de la variación evolutiva entre especies.** El genoma es una estructura dinámica que esta continuamente sujeta a las modificaciones causadas por los mecanismos evolutivos. Estos mecanismos, actuando a lo largo de millones de años, son los responsables de la secuencia de los genomas de los organismos que actualmente forman nuestra biosfera. Una comprensión profunda del funcionamiento del genoma solo es posible con un conocimiento paralelo de las diferencias de secuencias entre especies y de los procesos y mecanismos responsables de la aparición de estas diferencias a lo largo del tiempo.

Bibliografía

Estructura y composición de los genomas

BROWN, T. A: Genomes. 2nd ed.. Oxford, UK: BIOS Scientific Publishers Ltd; 2002. El texto completo de este libro de texto se puede consultar online gratuitamente en el NCBI Bookshelf (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Books).

LODISH, HARVEY; BERK, ARNOLD; ZIPURSKY, S. LAWRENCE; MATSUDAIRA, PAUL; BALTIMORE, DAVID; DARNELL, JAMES E: Molecular Cell Biology. 4th ed. New York: W. H. Freeman & Co.; c1999. También disponible online en el NCBI Bookshelf de forma gratuita.

Secuenciación. Artículos técnicos

SANGER, F., NICKLEN, S. & COULSON, A. R. DNA sequencing with chain-terminating inhibitors. Proc. Natl Acad. Sci. USA 74, 5463-5467 (1977).

SMITH, L. M. et al. Fluorescence detection in automated DNA sequence analysis. Nature 321, 674-679 (1986).

HUNKAPILLER, T., KAISER, R. J., KOOP, B. F. & HOOD, L. Large scale and automated DNA sequence determination. Science 254, 59-67 (1991).

MELDRUM, D. Automation for genomics. I. Preparation for sequencing. Genome Res. 10, 1081-1092 (2000).

MELDRUM, D. Automation for genomics. II. Sequencers, microarrays, and future trends. Genome Res. 10,1288-1303 (2000).

Métodos para la secuenciación a gran escala

Genome Analysis: A Laboratory Manual. 1. Analyzing DNA (eds BIRREN, B. et al.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997.

Genome Mapping and Sequencing. (ed IAN DUNHAM The Sanger Centre, Cambridge) Horizon Scientific Press. 2003.

Secuencia del genoma humano

Bases de datos con la Secuencia de Referencia y información exhaustiva sobre el genoma humano

Sanger Institute - http://www.ensembl.org/Homo_sapiens/
NCBI- http://www.ncbi.nlm.nih.gov/genome/guide/human/

Trabajos describiendo el borrador de la secuencia del genoma humano

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001).

VENTER, J. C. et al. The sequence of the human genome. Science 291, 1304-1351 (2001).

Secuencia de organismos modelo

Bacteria: *H. influenzae*

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496-512 (1995).

Levadura: *S. cerevisie*

GOFFEAU, A. et al. The yeast genome directory. Nature 387, S1-S105 (1997).

Nematodo: *C. elegans*

The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282, 2012-2018 (1998).

Mosca del vinagre: *D. melanogaster*

MYERS, E. W. et al. A whole-genome assembly of Drosophila. Science 287, 2196-2204 (2000).

Planta: *A. thaliana*

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796-815 (2000).

Raton: *M. musculus*

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62

Rata: *R. norvegicus*

Rat Genome Sequencing Consortium.Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature. 2004 Apr 1;428(6982):493-521.

Otros organismos

Un listado actualizado de los proyectos de secuenciación de organismos modelo se puede encontrar en : http://www.ncbi.nlm.nih.gov/Genomes/index.html