

El contenido del genoma

Microfotografía electrónica de barrido (MEB) de los cromosomas X e Y en metafase (35.000×). La mayoría del contenido del genoma se encuentra en el DNA cromosómico, aunque algunos orgánulos contienen sus propios genomas. © Biophoto Associates/Photo Researchers, Inc.

ÍNDICE

- 4.1** Introducción
- 4.2** Los genomas se pueden mapear en diversos niveles de resolución
- 4.3** Los genomas individuales muestran gran variación
- 4.4** Los RFLP y los SNP pueden utilizarse para el mapeo genético
- 4.5** ¿Por qué algunos genomas son tan grandes?
- 4.6** Los genomas eucariontes contienen tanto secuencias de DNA repetitivo como no repetitivo
- 4.7** Se pueden identificar genes eucariontes que codifican proteínas por la conservación de exones
- 4.8** La conservación de la organización del genoma contribuye a identificar genes
- 4.9** Algunos orgánulos tienen DNA
- 4.10** Los genomas de los orgánulos son DNA circulares que codifican las proteínas de los orgánulos
- 4.11** El genoma de cloroplasto codifica para muchas proteínas y para RNA
- 4.12** Los cloroplastos y las mitocondrias evolucionaron por endosimbiosis
- 4.13** Resumen

4.1 Introducción

Una pregunta clave acerca del genoma es cuántos genes contiene. Sin embargo, una pregunta más que fundamental es “¿qué es un gen?”. Claramente, los genes no se definen solamente como una secuencia de DNA que codifica un polipéptido, debido a que son muchos los genes que codifican múltiples polipéptidos, y a que muchos genes codifican RNA que sirven para otras funciones. Dadas la variedad de funciones del RNA y las complejidades de su expresión, parece prudente enfocarse en el concepto de gen como una “unidad de transcripción”. Sin embargo, grandes zonas de los cromosomas, que anteriormente se creía que carecían de genes, ahora parecen ser transcriptas ampliamente; por lo tanto, en la actualidad, es difícil definir a un “gen” con precisión.

Se puede intentar caracterizar tanto la cantidad total de genes como el número de genes que codifican proteínas en cuatro niveles, que corresponden a las etapas sucesivas de la expresión génica:

- ▶ **genoma** Grupo completo de secuencias en el material genético de un organismo. Incluye la secuencia de cada cromosoma más cualquier DNA presente en los orgánulos.
 - ▶ **transcriptoma** Conjunto completo de los RNA presente en una célula, tejido u organismo. Su complejidad se debe principalmente al mRNA, pero también incluye RNA no codificantes.
 - ▶ **proteoma** Grupo completo de proteínas que se expresan en el genoma. A veces, el término se utiliza para describir el grupo de proteínas expresadas por una célula en un momento dado.
 - ▶ **interactoma** Grupo completo de complejos proteicos o de interacciones entre proteínas presentes en una célula, tejido u organismo.
- El **genoma** es el grupo completo de genes de un organismo. En última instancia, se define por la secuencia de DNA completa, aunque de manera práctica no es posible identificar cada gen de manera inequívoca solamente sobre la base de la secuencia.
 - El **transcriptoma** es el grupo completo de genes expresados bajo condiciones particulares. Se define en términos del grupo de moléculas de RNA que están presentes y pueden referirse a un tipo celular único, o a cualquier organización compleja de células, hasta constituir un organismo. Debido a que algunos genes generan múltiples mRNA, es probable que el transcriptoma sea más grande que el número de genes definidos directamente en el genoma. El transcriptoma incluye los RNA no codificantes (como tRNA, rRNA, y microRNA o miRNA, que se describen en la Sección 13.8, *Los eucariontes contienen RNA reguladores*) así como también mRNA.
 - El **proteoma** es el grupo de polipéptidos codificados por el genoma entero o producidos en cualquier célula o tejido particular. Se corresponde con los mRNA en el transcriptoma, aunque existen diferencias de detalles que reflejan cambios en la abundancia relativa o en las estabildades de los mRNA y de las proteínas. También pueden existir modificaciones postraduccionales de las proteínas que permiten que se produzcan más de una proteína a partir de un único transcripto (esto se denomina *corte y empalme de proteína*; véase la Sección 29.11, *El corte y empalme de proteínas es autocatalítico*).
 - Las proteínas pueden funcionar de forma independiente o como parte de complejos multiproteicos o multimoleculares, como las holoenzimas y las vías metabólicas donde se agrupan las enzimas. Dos ejemplos son la holoenzima RNA polimerasa (véase la Sección 11.5, *La RNA polimerasa bacteriana consiste en la enzima central y el factor sigma*) y el empalmosoma (véase la Sección 28.8, *Cinco snRNP forman el empalmosoma*). Si se pudieran identificar todas las interacciones proteína-proteína, se podría definir el número total de complejos independientes de proteínas. Esto a menudo se denomina como el **interactoma**.

El número máximo de genes que codifican proteínas en el genoma puede identificarse directamente mediante la caracterización de marcos de lectura abiertos. El mapeo a gran escala de esta naturaleza es complicado por el hecho de que los genes interrumpidos pueden consistir en muchos marcos de lectura abiertos separados. No necesariamente se tiene información acerca de las funciones de estos productos proteicos –o, en efecto, la prueba de que son todos expresados–; por lo tanto, este enfoque se restringe a definir el *potencial* del genoma. Sin embargo, existe una suposición fuerte de que cualquier marco de lectura abierto conservado es probable que sea expresado.

Otro enfoque consiste en definir la cantidad de genes directamente en términos del transcriptoma (por identificación directa de todos los RNA) o del proteoma (por identificación directa de todos los polipéptidos). Esto brinda una certidumbre de que se está tratando con genes *bona fide* que son expresados bajo circunstancias conocidas. Esto permite preguntarse cuántos genes se expresan en un tejido o tipo celular particular, qué variaciones existen en los niveles relativos de expresión, y cuántos de los genes expresados en una célula particular son únicos de esa célula o se expresan además en otras. Además, el transcriptoma puede revelar cuántos mRNA diferentes (p. ej., los mRNA que contienen diferentes combinaciones de exones) se generan a partir de un gen dado.

En relación con los tipos de genes, se puede preguntar si un gen particular es *esencial*: ¿qué le sucede a un mutante nulo? Si una mutación nula es letal, o si el organismo tiene un defecto visible, se puede concluir que el gen es esencial o, al menos, que porta una ventaja selectiva. Sin embargo, algunos genes pueden eliminarse sin efecto aparente en el fenotipo. ¿Son estos genes realmente prescindibles, o se produce una desventaja selectiva a partir de la ausencia del gen, –tal vez en otras circunstancias, o durante períodos más largos? En algunos casos, la ausencia de estos genes podría ser compensada por un mecanismo redundante, como una duplicación génica, que proporciona una copia de resguardo para una función esencial.

REVISIÓN DE CONCEPTOS

Explicar por qué, en los animales el transcriptoma y el proteoma de una célula suelen ser más pequeños que su genoma; mientras que el transcriptoma y el proteoma del organismo entero, en general, son más grandes que su genoma.

4.2 Los genomas se pueden mapear en diversos niveles de resolución

Definir los contenidos de un genoma esencialmente significa “trazar un mapa”. Se puede pensar acerca del mapeo de genes y genomas en varios niveles de resolución:

- Un **mapa genético** (o **de ligamiento**) identifica la distancia entre los loci en términos de sus frecuencias de recombinación. Está limitado, ya que depende de la ocurrencia de recombinación de diferentes marcadores que puedan ser visibles –como un rasgo del fenotipo– o que puedan visualizarse –mediante electroforesis. Por ejemplo, se puede construir un **mapa de ligamiento** por medición de la recombinación entre sitios en el DNA genómico que tienen variaciones de secuencias que generan diferencias en la susceptibilidad a la escisión por parte de ciertas enzimas de restricción. Debido a que tales variaciones son comunes, se puede preparar un mapa para cualquier organismo independientemente de la aparición de mutantes. Debido a que, por la distancia física entre los sitios, las frecuencias de recombinación pueden estar distorsionadas, un mapa de ligamiento no representa de forma precisa las distancias físicas entre el material genético.
- Un **mapa de restricción** se construye por ruptura del DNA en fragmentos con enzimas de restricción y medición de las distancias físicas, en términos de la longitud del DNA (determinada por migración en un gel por electroforesis), entre los sitios de corte. Un mapa de restricción no identifica intrínsecamente los sitios de interés, como por ejemplo un gen. Para que sea relativo a un mapa genético, se deben caracterizar las mutaciones en términos de sus efectos sobre los sitios de restricción. Se pueden reconocer los cambios grandes en el genoma debido a que afectan los tamaños o cantidades de los fragmentos de restricción. Las mutaciones puntuales son más difíciles de detectar debido a que sólo cambia un sitio de restricción único, o porque se encuentran entre sitios de restricción y son indetectables.
- El mapa genómico final es la secuencia de DNA. A partir de la secuencia, se puede identificar los genes y las distancias entre ellos. Mediante el análisis del potencial de codificar una proteína que tiene una secuencia de DNA, se pueden formular hipótesis acerca de su función. La presunción básica aquí es que la selección natural evita la acumulación de mutaciones dañinas en las secuencias que codifican proteínas. Revirtiendo este argumento, se puede asumir que una secuencia codificante intacta es probable que sea utilizada para generar una proteína.

Al comparar una secuencia de DNA de tipo silvestre con la de un alelo mutante, se puede determinar la naturaleza de una mutación y el sitio exacto de aparición. Esto proporciona una manera de determinar la relación entre el mapa genético (basado completamente en los sitios de mutación) y el mapa físico (basado en la secuencia de DNA).

Se han utilizado técnicas similares para identificar y secuenciar genes y mapear el genoma, aunque por supuesto existe una diferencia de escala. En cada caso, el principio es

► **mapa genético** Véase mapa de ligamiento

► **mapa de ligamiento** Mapa de las posiciones de los loci o de otros marcadores genéticos sobre un cromosoma obtenido por medición de las frecuencias de recombinación entre marcadores.

► **mapa de restricción** Orden lineal de sitios de restricción en el DNA, el cual es determinado mediante cortes en la molécula con varias endonucleasas de restricción, o mediante la búsqueda de los sitios de restricción en una secuencia conocida.

para caracterizar una serie de fragmentos superpuestos de DNA que pueden conectarse dentro de un mapa continuo. La característica crucial es que, en el mapa, cada segmento se relaciona con el segmento siguiente mediante el solapamiento entre ellos, de manera tal de poder asegurar que no se pierdan segmentos. Este principio se aplica tanto en el ordenamiento de los fragmentos grandes en un mapa como en la conexión de las secuencias que forman los fragmentos.

CONCEPTOS CLAVE

- Los mapas de ligamiento se basan en la frecuencia de recombinación entre los marcadores genéticos; los mapas de restricción se basan en las distancias físicas entre los marcadores.
- Se puede utilizar la caracterización molecular de las mutaciones para compaginar los mapas de ligamiento con los mapas físicos.

REVISIÓN DE CONCEPTOS

Si se toma una muestra del mismo cromosoma de poblaciones diferentes del mismo organismo, el mapa físico puede ser idéntico; pero el mapa de ligamiento puede ser levemente diferente. ¿Por qué?

4.3 Los genomas individuales muestran gran variación

El punto de vista mendeliano del genoma clasificaba los alelos según dos alternativas: de tipo silvestre o mutante. Posteriormente, se reconoció la existencia de alelos múltiples, cada uno con diferente efecto sobre el fenotipo. En algunos casos, puede incluso no ser adecuado definir ningún alelo como “de tipo silvestre”.

► **polimorfismo** Aparición simultánea en la población de alelos que muestran variaciones en una posición determinada.

La coexistencia de alelos múltiples en un locus se denomina **polimorfismo** genético. Cualquier sitio en el cual existan alelos múltiples como componentes estables de la población es, por definición, polimórfico. Un locus suele definirse como polimórfico si dos o más alelos están presentes a una frecuencia > 1% en la población.

Aunque no se evidencia a partir del fenotipo, el tipo silvestre puede, en sí mismo, ser polimórfico. Se pueden distinguir versiones múltiples del alelo de tipo silvestre por las diferencias en la secuencia que no afectan a su función y que, por lo tanto, no producen variantes en los fenotipos. Una población puede ser ampliamente polimórfica desde el punto de vista del genotipo. Muchas variantes de secuencias diferentes pueden existir en un locus dado; algunas de ellas son evidentes debido a que afectan al fenotipo, pero otras se encuentran ocultas debido a que no tienen efecto visible.

Por lo tanto, puede haber un continuo de cambios en un locus, incluso aquellos que cambian la secuencia de DNA –pero que no cambian la secuencia de proteína–, en aquellos que cambian la secuencia de proteína –sin cambiar su función–, en aquellos que crean proteínas con actividades diferentes, y en aquellos que crean proteínas mutantes que son no funcionales.

► **polimorfismo de un nucleótido único (SNP)** Polimorfismo (variación en la secuencia entre individuos) causado por un cambio en un solo nucleótido. Responsable de la mayoría de la variación genética entre individuos.

Un cambio en un nucleótido cuando se comparan los alelos se denomina **polimorfismo de un nucleótido único (SNP–single nucleotide polymorphism)**. En promedio, se produce uno cada ~1.330 bases en el genoma humano. Definido por sus SNP, cada ser humano es único. Los SNP pueden detectarse a través de diferentes maneras, que abarcan desde la comparación directa de secuencias hasta la espectroscopia de masa o los métodos bioquímicos que producen diferencias basadas en variaciones de la secuencia en una región definida.

Una meta del mapa genético consistió en obtener un catálogo de las variantes comunes. La frecuencia observada de los SNP por genoma pronosticó que, en la población humana en su totalidad (considerando la suma de todos los genomas humanos de todos los individuos vivos), debería haber > 10 millones de SNP que aparecerían a una frecuencia > 1%. Ya se han identificado > 1 millón.

Algunos polimorfismos en el genoma pueden detectarse al comparar los mapas de restricción de diferentes individuos. El criterio es un cambio en el patrón de fragmentos de DNA que se produce por una escisión con una enzima de restricción. La **FIGURA 4.1** muestra que, cuando un sitio diana está presente en el genoma de un individuo y ausente

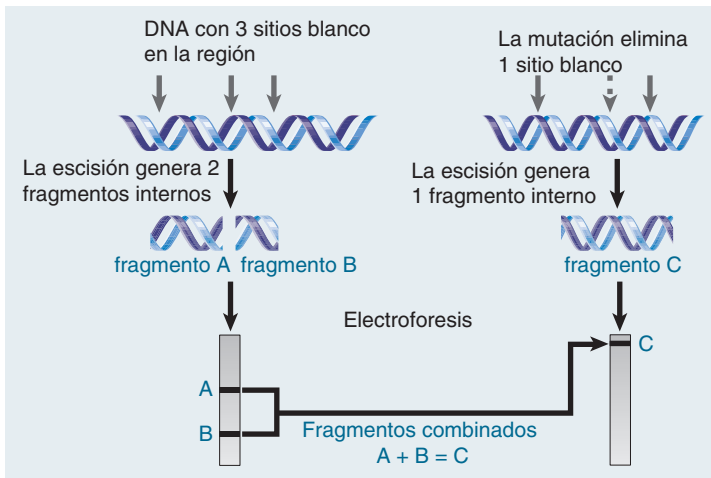


FIGURA 4.1 Una mutación puntual, que afecta un sitio de restricción, es detectada por una diferencia en los fragmentos de restricción.

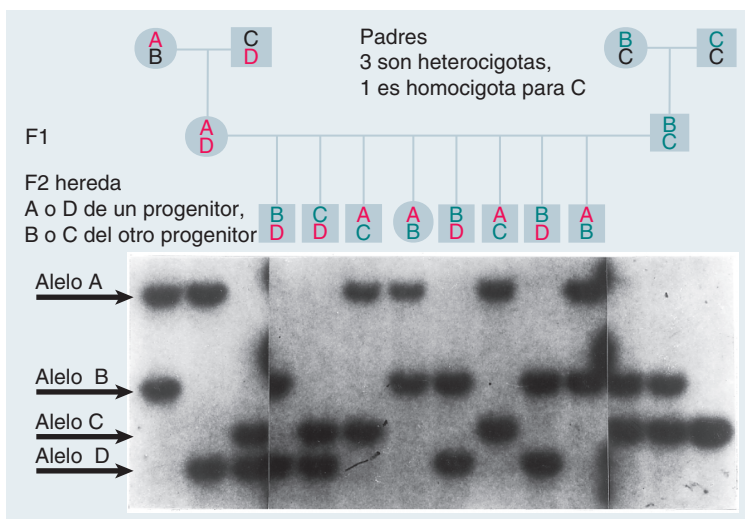


FIGURA 4.2 Los polimorfismos en los sitios de restricción se heredan de acuerdo con las reglas mendelianas. Mediante la técnica de transferencia de Southern y la hibridación con sondas, se visualizaron cuatro alelos para un marcador de restricción en todas las combinaciones de a pares posibles y segregaron de manera independiente en cada generación. Fotografía cortesía de Ray White, Ernest Gallo Clinic and Research Center, University of California, San Francisco.

en otros, la escisión extra en el primer genoma generará dos fragmentos correspondientes al fragmento único en el segundo genoma. Una diferencia en los mapas de restricción entre dos individuos se denomina un **polimorfismo de longitud de los fragmentos de restricción (RFLP—restriction fragment length polymorphism)**. Básicamente, un RFLP es un SNP que se localiza en el sitio diana de una enzima de restricción. Puede ser utilizado como un marcador genético exactamente de la misma forma que cualquier otro marcador. En lugar de examinar algunas características del fenotipo, directamente se evalúa el genotipo, como lo revela el mapa de restricción. La **FIGURA 4.2** muestra un pedigrí de un polimorfismo de restricción seguido a través de tres generaciones. Ésta muestra la segregación mendeliana de los fragmentos de DNA marcadores.

El mapa de restricción es independiente de la función génica: en consecuencia, un RFLP en este nivel puede detectarse independientemente si el cambio de secuencia afecta al fenotipo. Probablemente muy pocos de los RFLP en un genoma realmente afecten al fenotipo. La mayoría de los cambios de las secuencias involucradas no tienen efecto sobre la producción de proteínas (p. ej., porque éstas se sitúan entre los genes).

CONCEPTOS CLAVE

- El polimorfismo puede detectarse en distintos niveles: como fenotipo cuando una secuencia afecta la función génica; en el fragmento de restricción, cuando afecta al sitio diana de la enzima de restricción; y en la secuencia, por análisis directo del DNA.
- Los alelos de un gen muestran gran polimorfismo en la secuencia, pero muchos cambios de secuencia no afectan la función.

► **polimorfismo de longitud de los fragmentos de restricción (RFLP)** Diferencias heredadas en los sitios para las enzimas de restricción —por ejemplo, por cambios de bases en el sitio blanco—, las cuales generan diferencias en las longitudes de los fragmentos producidos por el corte con una enzima determinada. Se utilizan para el mapa genético y para vincular al genoma directamente con un marcador genético convencional.

REVISIÓN DE CONCEPTOS

¿Por qué una mutación en la secuencia codificante podría producir un polimorfismo genético, pero no uno en el fenotipo?

4.4 Los RFLP y los SNP pueden utilizarse para el mapeo genético

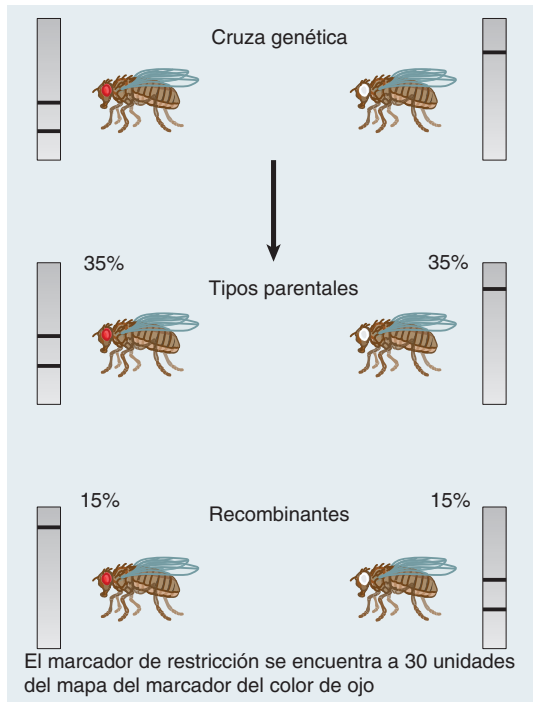


FIGURA 4.3 Se puede utilizar un polimorfismo de restricción como marcador genético para medir la distancia de recombinación a partir de un marcador fenotípico (como el color de ojos). La figura simplifica la situación al mostrar sólo las bandas de DNA que corresponden al alelo de un genoma en un diploide.

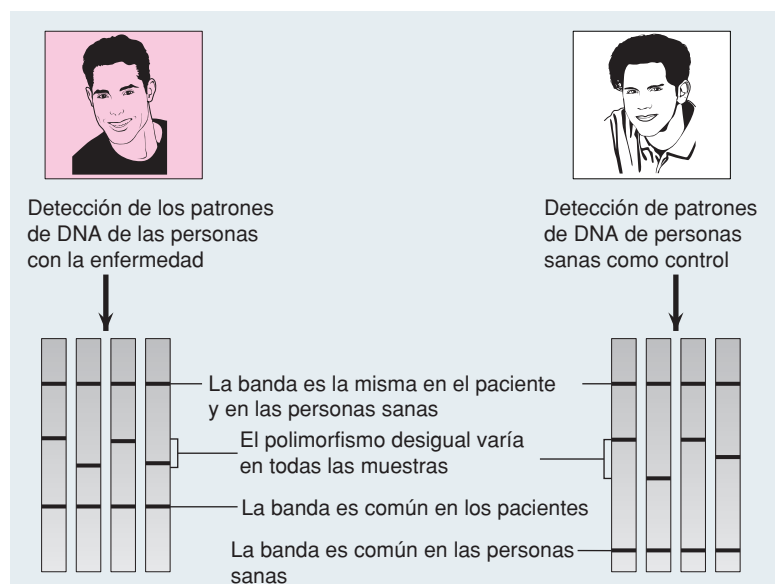
FIGURA 4.4 Si un marcador de restricción se asocia con una característica fenotípica, el sitio de restricción debe localizarse cerca del gen responsable para el fenotipo. La mutación que cambia la banda que es común en personas sanas por la banda que es común en pacientes, está estrechamente ligada al gen de la enfermedad.

Se puede medir la frecuencia de recombinación entre un marcador de restricción y un marcador con fenotipo visible, como se ilustra en la **FIGURA 4.3**. Por ende, un mapa genético puede incluir tanto marcadores de genotipos como de fenotipos.

Los marcadores de restricción no se limitan a los cambios en el genoma que afectan al fenotipo; como consecuencia, los marcadores proporcionan la base de una técnica extremadamente poderosa para identificar las variantes genéticas a nivel molecular. Un problema típico atañe a una mutación con efectos conocidos sobre el fenotipo, donde el locus genético relevante puede colocarse en un mapa genético, pero del cual no se tiene conocimiento acerca del gen o de la proteína correspondiente. Muchas enfermedades humanas severas o mortales se encuentran en esta categoría. Por ejemplo, la fibrosis quística exhibe una herencia mendeliana, pero la naturaleza molecular de la función mutante fue desconocida hasta que se la pudo identificar como consecuencia de la caracterización del gen.

Si los polimorfismos de restricción se producen de manera aleatoria en el genoma, deberían existir algunos cerca de cualquier gen diana particular. Se pueden identificar esos marcadores de restricción en virtud de sus asociaciones estrechas con el fenotipo mutante. Si se compara el mapa de restricción del DNA de pacientes que sufren una enfermedad con el DNA de personas sanas, se podría encontrar que un sitio de restricción particular siempre está presente (o siempre ausente) en los pacientes.

En la **FIGURA 4.4**, se muestra un ejemplo hipotético. Esta situación corresponde al hallazgo del 100% de ligamiento entre el marcador de restricción y el locus que produce el fenotipo. Esto podría implicar que el marcador de restricción se encuentra tan cerca del gen mutante que nunca son separados por recombinación; de hecho, pueden ser la misma mutación.



La identificación de tal marcador tiene dos consecuencias importantes:

- Puede ofrecer un procedimiento de diagnóstico para la detección de la enfermedad. Algunas de las enfermedades humanas conocidas tienen un patrón heredable, pero la enfermedad que se define en términos moleculares no puede ser diagnosticada fácilmente. Si un marcador de restricción está estrechamente ligado al fenotipo, entonces su presencia puede utilizarse como diagnóstico de probabilidad de portar el alelo de la enfermedad.
- Puede conducir al aislamiento del gen. El marcador de restricción debe encontrarse relativamente cerca del gen en el mapa genético si los dos loci raramente o nunca se recombinan. En términos genéticos, “relativamente cerca” puede ser una distancia sustancial en términos de pares de bases del DNA; no obstante, proporciona un punto de partida a partir del cual se puede proceder a lo largo del DNA hasta el gen mismo.

La frecuencia de aparición de los SNP en el genoma humano los torna útiles para el mapeo genético. De los $1,5 \times 10^6$ SNP que ya se han identificado, existe en promedio un SNP cada 1 a 2 kb. Esto debería permitir la localización rápida de nuevos genes asociados a una enfermedad por la localización de estos mismos entre los SNP más cercanos.

Sobre la base del mismo principio, el mapeo de los RFLP se ha estado utilizando durante algún tiempo. Una vez que un RFLP se ha asignado a un grupo de ligamiento (es decir, a un cromosoma), puede ser colocado en un mapa genético. El mapeo de RFLP, tanto en los genomas humano como en los del ratón ha conducido a la construcción de mapas de ligamiento para ambos. Cualquier sitio con una posición desconocida puede ser evaluado para ligamiento a estos sitios y, por este principio, puede ser rápidamente colocado en el mapa. Existen menos RFLP que SNP, de manera tal que la resolución del mapa de RFLP es, en principio, más limitada.

La gran proporción de sitios polimórficos significa que cada individuo tiene una constelación única de SNP y RFLP. La combinación particular de sitios que se encuentran en una región específica se denomina **haplotipo**, un genotipo en miniatura. El haplotipo fue introducido inicialmente como un concepto para describir la constitución genética del locus del complejo mayor de histocompatibilidad: una región que codifica proteínas de importancia en el sistema inmunitario (véase el Capítulo 22, *Diversidad inmunitaria*). En la actualidad, el término se ha extendido para describir la combinación particular de alelos, los sitios de restricción o de cualquier otro marcador genético presente en algunas zonas definidas del genoma. Mediante el uso de los SNP, se realiza un mapa de haplotipo detallado del genoma humano; esto permite que los genes causantes de enfermedad sean localizados más fácilmente en el mapa.

La existencia de los RFLP proporciona las bases de una técnica para establecer de manera inequívoca las relaciones entre progenitores y descendencia. En casos de duda de parentesco, una comparación del mapa de RFLP en un cromosoma apropiado entre padres y niños potenciales permite la asignación absoluta del parentesco. El uso del análisis de restricción del DNA para identificar individuos se denomina **huella genética**. El análisis de las secuencias “minisatélites” especialmente variables se utiliza para realizar el mapeo del genoma humano (véase la Sección 6.12, *Los minisatélites son útiles para el mapeo de genes*).

CONCEPTOS CLAVE

- Los RFLP y los SNP pueden ser la base para los mapas de ligamiento y son útiles para establecer las relaciones parentales.

REVISIÓN DE CONCEPTOS

Describir cómo la posición de un alelo de enfermedad puede ser delimitada a una región cromosómica específica mediante el análisis de ligamiento utilizando los RFLP de miembros de una familia grande en la cual el alelo de la enfermedad esté segregado.

- **haplotipo** Combinación particular de alelos en una región definida de un cromosoma de la que resulta un genotipo en miniatura. Anteriormente se lo utilizaba para describir las combinaciones de alelos del complejo mayor de histocompatibilidad (CMH); actualmente puede usarse para describir combinaciones determinadas de RFLP, SNP u otros marcadores.
- **huella genética** Técnica para analizar las diferencias entre individuos con los fragmentos generados con enzimas de restricción que cortan regiones que contienen secuencias cortas repetidas, o por PCR. Las longitudes de las regiones repetidas son únicas para cada individuo, y, por ello, la presencia de un subgrupo determinado en dos individuos puede utilizarse para determinar su herencia común (p. ej., la relación padre-hijo).

4.5 ¿Por qué algunos genomas son tan grandes?

La cantidad total de DNA en el genoma (haploide) es una característica de cada especie viviente que se conoce como **valor C**. Existe una variación enorme en el intervalo de

- **valor C** Cantidad total de DNA del genoma (por conjunto haploide de cromosomas).

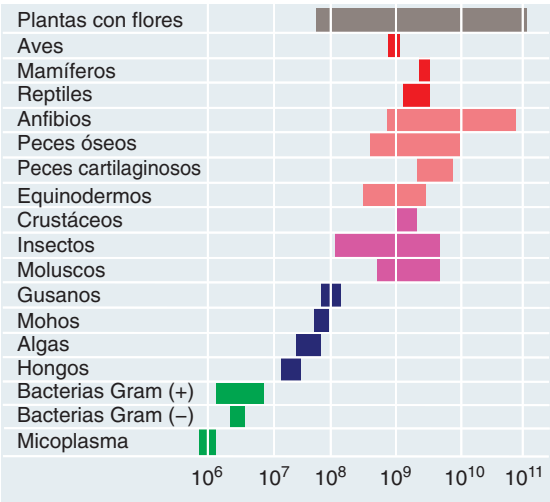


FIGURA 4.5 El contenido de DNA del genoma haploide se incrementa con la complejidad morfológica de los eucariontes inferiores, pero varía en gran medida dentro de algunos grupos de eucariontes superiores. El intervalo de valores de DNA dentro de cada grupo se indica mediante el área sombreada.

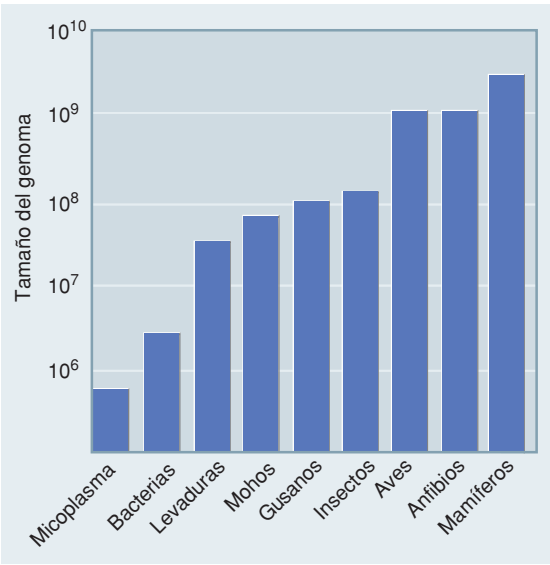


FIGURA 4.6 El tamaño del genoma mínimo que se encuentra en cada taxón aumenta desde los procariontes hasta los mamíferos.

| Filo | Especie | Genoma (pb) |
|-----------------|-------------------------|-------------------|
| Alga | <i>Pyrenomas salina</i> | $6,6 \times 10^5$ |
| Micoplasma | <i>M. pneumoniae</i> | $1,0 \times 10^6$ |
| Bacteria | <i>E. coli</i> | $4,2 \times 10^6$ |
| Levadura | <i>S. cerevisiae</i> | $1,3 \times 10^7$ |
| Moho deslizante | <i>D. discoideum</i> | $5,4 \times 10^7$ |
| Nemátodo | <i>C. Elegans</i> | $8,0 \times 10^7$ |
| Insecto | <i>D. Melanogaster</i> | $1,8 \times 10^8$ |
| Ave | <i>G. Domesticus</i> | $1,2 \times 10^9$ |
| Anfibio | <i>X. Laevis</i> | $3,1 \times 10^9$ |
| Mamífero | <i>H. sapiens</i> | $3,3 \times 10^9$ |

FIGURA 4.7 Tamaños de genomas de algunos organismos experimentales comunes.

los valores C, desde $< 10^6$ pb para un micoplasma hasta $> 10^{11}$ pb para algunas plantas y anfibios.

La **FIGURA 4.5** resume el intervalo de los valores C que se encuentran en diferentes taxones (grupos de organismos clasificados juntos). Existe un incremento en el tamaño de genoma mínimo que se encuentra en cada grupo a medida que la complejidad se acrecienta. Aunque los valores C son superiores en los eucariontes multicelulares, se pueden observar variaciones amplias en los tamaños de genomas dentro de algunos taxones.

Como se observa en la **FIGURA 4.6**, graficar la cantidad mínima de DNA requerida para un miembro de cada grupo sugiere que es necesario un incremento en el tamaño del genoma para formar procariontes más complejos y eucariontes inferiores.

Los micoplasmas son los procariontes más pequeños y tienen genomas de sólo ~3 veces el tamaño de un bacteriófago grande. Los tamaños de genomas bacterianos más típicos comienzan a partir de $\sim 2 \times 10^6$ pb. Los eucariontes unicelulares (cuyos estilos de vida pueden asemejarse al de los procariontes) se les arreglan con genomas que también son pequeños, aunque son más grandes que los de las bacterias. Ser eucariente *per se*, no implica un gran incremento en el tamaño del genoma; una levadura puede tener un tamaño de genoma de $\sim 1,3 \times 10^7$ pb, que sólo es alrededor del doble de tamaño de un genoma bacteriano promedio.

Una duplicación adicional del tamaño del genoma es adecuada para sustentar al moho deslizante *Dictyostelium discoideum*, que es capaz de vivir de modo unicelular o multicelular. Se necesita otro incremento en la complejidad para producir los primeros organismos completamente multicelulares; el gusano nemátodo *Caenorhabditis elegans* tiene un contenido de DNA de 8×10^7 pb.

También se puede observar un incremento estable en el tamaño del genoma con la complejidad en la lista de la **FIGURA 4.7** de algunos organismos comúnmente analizados. Es necesario incrementar el tamaño del genoma para formar insectos, aves o anfibios, y mamíferos. Si bien, después de este punto, no es clara la relación entre el tamaño del genoma y la complejidad morfológica del organismo.

Se sabe que los genes son mucho más grandes que las secuencias necesarias para codificar los polipéptidos, debido a que los exones pueden abarcar sólo una parte pequeña de la longitud total de un gen. Esto explica por qué existe mucho más DNA del que es necesario para proporcionar marcos de lectura para todas las proteínas del organismo. Grandes regiones de un gen interrumpido pueden no codificar un polipéptido. Además, también puede haber longitudes significativas de DNA entre los genes. De manera tal que no es posible deducir el número de genes a partir del tamaño global del genoma.

La **paradoja del valor C** se refiere a la falta de correlación entre el tamaño del genoma y la complejidad genética. También existen casos curiosos acerca del tamaño relativo del genoma; por ejemplo, el batracio *Xenopus* y los seres humanos tienen genomas, esencialmente, del mismo tamaño. En algunos taxones, existen variaciones extremadamente grandes en el contenido de DNA entre los organismos que no varían mucho en complejidad (véase la Figura 4.5). (Esto es especialmente notable en los insectos, anfibios y plantas, pero no se produce en las aves, reptiles ni mamíferos, que tienen pequeñas variaciones dentro del grupo, una variación de ~2 veces el rango del tamaño del genoma.) El grillo tiene un genoma de 11 veces, el tamaño del genoma de la mosca de la fruta. En los anfibios, los genomas más pequeños son $< 10^9$ pb, mientras que los más grandes son $\sim 10^{11}$ pb. Es improbable que exista una diferencia grande en la cantidad de genes necesarios para especificar estos anfibios. Todavía no se comprende del todo hasta qué punto esta variación es selectivamente neutral o está sujeta a la selección natural.

CONCEPTOS CLAVE

- No existe correlación entre el tamaño del genoma y la complejidad genética.
- Existe un incremento en el tamaño mínimo del genoma necesario para constituir organismos de complejidad creciente.
- Existen amplias variaciones en los tamaños del genoma de los organismos dentro de muchos taxones.

REVISIÓN DE CONCEPTOS

¿Cuáles son algunas de las ventajas y desventajas de un genoma grande? ¿Y de un genoma pequeño?

► **paradoja del valor C** Falta de relación entre el contenido de DNA (valor C) de un organismo y su potencial de codificación.

4.6 Los genomas eucariontes contienen tanto secuencias de DNA repetitivo como no repetitivo

En general, la naturaleza del genoma eucarionte puede ser evaluada según la cinética de reasociación y desnaturalización del DNA. Esta técnica se utilizó ampliamente antes de que fuera posible la secuenciación del DNA a gran escala.

Las cinéticas de reasociación identifican dos tipos generales de secuencias genómicas:

- El **DNA no repetitivo** consiste en secuencias que son únicas: sólo existe una copia en un genoma haploide.
- El **DNA repetitivo** consiste en secuencias que están presentes en más de una copia en cada genoma.

A menudo el DNA repetitivo se divide en dos tipos generales:

- El DNA moderadamente repetitivo consiste en secuencias relativamente cortas que se repiten en general entre 10-1.000 veces en el genoma. Las secuencias están dispersadas a través de todo el genoma y son responsables del alto grado de formación de estructuras secundarias en el pre-mRNA, cuando las repeticiones invertidas en los intrones se aparean para formar regiones dúplex.
- El DNA altamente repetitivo consiste en secuencias muy cortas (en general < 100 pb) que están presentes muchos cientos de veces en el genoma, suele estar organizado como largas regiones de repeticiones en tándem (véase la Sección 6.9, *Los DNA satélites suelen encontrarse en la heterocromatina*). Ninguna clase se encuentra en los exones.

La proporción del genoma ocupada por el DNA no repetitivo varía ampliamente entre los taxones. La **FIGURA 4.8** resume la organización del genoma de algunos organismos representativos. Los procariotes contienen principalmente DNA no repetitivo. Para los eucariontes inferiores, la mayoría del DNA es no repetitivo; < 20% se encuentra en la categoría de uno o más componentes moderadamente repetitivos. En las células animales, hasta la mitad del DNA suele estar ocupado por componentes moderada y altamente repetitivos. En las plantas y en los anfibios, los componentes moderada y altamente repetitivos pueden representar hasta el 80% del genoma, de manera tal que el DNA no repetitivo se reduce a un componente minoritario.

Una parte significativa del DNA moderadamente repetitivo son los **transposones (elementos transponibles)**: consisten en secuencias cortas de DNA (~1 kb) que tienen la capacidad de moverse a localizaciones nuevas en el genoma y/o de hacer copias adicionales de sí mismos (véase el Capítulo 21, *Transposones, retrovirus, y retrotransposones*). En algunos genomas de eucariontes superiores, pueden incluso ocupar más de la mitad del genoma (véase la Sección 5.5, *El genoma humano tiene menos genes de lo esperado originalmente*).

Los transposones suelen ser vistos como **DNA redundante**, que se define como secuencias que se propagan a sí mismas dentro de un genoma sin contribuir al desarrollo ni al funcionamiento del organismo. Los transposones pueden causar reordenamientos del genoma,

► **DNA no repetitivo** DNA que es único (presente sólo una vez) en un genoma.

► **DNA repetitivo** DNA que está presente en muchas copias (relacionadas o idénticas) en un genoma.

► **Transposones (elementos transponibles)** Secuencia de DNA capaz de insertarse a sí misma (o una copia de sí misma) a una nueva ubicación en el genoma sin tener ninguna relación en cuanto a secuencia con el locus diana.

► **DNA redundante** Secuencias de DNA que no contribuyen al fenotipo del organismo, sino que tienen la autoperpetuación dentro del genoma como única función.

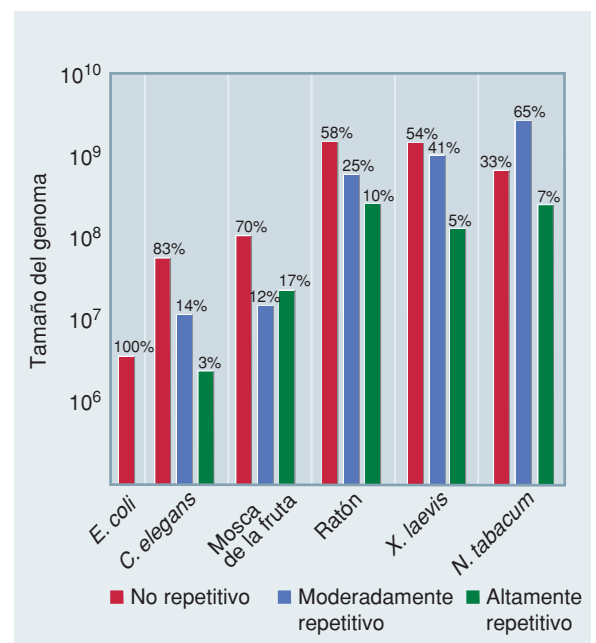


FIGURA 4.8 Las proporciones de los diferentes componentes de secuencia varían en los genomas de eucariontes. El contenido absoluto de DNA no repetitivo se incrementa con el tamaño del genoma, pero alcanza una meseta en $\sim 2 \times 10^9$ pb.

y podrían conferir ventajas selectivas. Es justo decir, no obstante, que no se comprende realmente por qué las fuerzas selectivas no actúan contra los transposones convirtiéndolos en una porción grande del genoma. Puede ser que éstos sean selectivamente neutrales en la medida en que no interrumpen o eliminan regiones codificantes o reguladoras. Sin embargo, muchos organismos suprimen activamente la transposición, como en algunos casos en los que se producen rupturas cromosómicas deletéreas (véase la Figura 21.18). Otro término que se utiliza para describir el exceso aparente del DNA en algunos genomas es el DNA “basura”, lo que implica secuencias genómicas sin función aparente. Por supuesto, es probable que exista un equilibrio en el genoma entre la generación de secuencias nuevas y la eliminación de secuencias no deseadas, y alguna proporción del DNA que aparentemente carece de función puede estar en proceso de ser eliminada.

La longitud del componente del DNA no repetitivo tiende a incrementarse con el tamaño global del genoma a medida que se avanza hasta un tamaño total de genoma de $\sim 3 \times 10^9$ (característico de los mamíferos). Incrementos posteriores en el tamaño del genoma, sin embargo, suelen reflejar un aumento en la cantidad y proporción de los componentes repetitivos, de manera que es poco común que un organismo presente un componente de DNA no repetitivo $> 2 \times 10^9$. Por lo tanto, el contenido de DNA no repetitivo de los genomas concuerda de mejor manera con el sentido que se tiene de la complejidad relativa de los organismos. *E. coli* tiene $4,2 \times 10^6$ pb de DNA no repetitivo, *C. elegans* tiene un orden más de magnitud ($6,6 \times 10^7$ pb), *D. melanogaster* tiene $\sim 10^8$ pb, y los mamíferos aún otro orden de magnitud más, $\sim 2 \times 10^9$ pb.

¿Qué tipo de DNA corresponde a los genes que codifican polipéptidos? Las cinéticas de reasociación en general muestran que el mRNA deriva del DNA no repetitivo. Por lo tanto, la cantidad de DNA no repetitivo es un mejor indicador del potencial codificante de lo que es el valor C. (Sin embargo, un análisis más detallado sobre la base de las secuencias genómicas demostró que muchos exones tienen secuencias que están relacionadas con otros exones [véase la Sección 3.4, *Las secuencias de los exones se conservan, pero los intrones varían*]. Tales exones evolucionan a partir de una duplicación para dar copias que, inicialmente, son idénticas, pero que luego divergieron en su secuencia durante la evolución.)

CONCEPTOS CLAVE

- Las cinéticas de reasociación del DNA, después de que un genoma ha sido desnaturalizado, distinguen secuencias por su frecuencia de repetición en el genoma.
- Los polipéptidos están generalmente codificados por secuencias en el DNA no repetitivo.
- Los genomas más grandes dentro de un taxón no contienen más genes, sino que tienen cantidades grandes de DNA repetitivo.
- Una gran parte del DNA moderadamente repetitivo puede estar constituida por transposones.

REVISIÓN DE CONCEPTOS

¿Qué es la resolución de la paradoja del valor C?

4.7 Se pueden identificar genes eucariontes que codifican proteínas por la conservación de exones

Algunos enfoques importantes para identificar genes eucariontes que codifican proteínas se basan en el contraste entre la conservación de exones y la variación de intrones. En una región que contiene un gen cuya función ha sido conservada entre un espectro de especies, la secuencia que representa al polipéptido debería tener dos propiedades distintas:

- debe tener un marco de lectura abierto, y
- es probable que posea una secuencia (ortóloga) relacionada, en otras especies.

Estas características pueden utilizarse para aislar genes.

Suponiendo que se conozca por análisis de ligamiento que un gen, que influye en un rasgo particular, se localiza en una región cromosómica dada; si se carece de conocimiento acerca de la naturaleza del producto génico, ¿cómo se hace para identificar el gen en una región que puede ser de un tamaño, por ejemplo, de > 1 Mb?

Una técnica exitosa realizada con algunos genes de importancia médica consistió en examinar fragmentos relativamente cortos de la región para las dos propiedades esperadas de un gen conservado. Primero, se buscó identificar fragmentos con hibridación cruzada con los genomas de otras especies, y luego se analizaron esos fragmentos en busca de marcos de lectura abiertos.

El primer criterio se aplica mediante una **transferencia zoo (zoo blot)**. Se utilizan fragmentos cortos de la región como sondas marcadas para evaluar los DNA homólogos de una variedad de especies mediante la hibridación Southern (una técnica para transferir fragmentos de DNA separados por electroforesis en gel a una membrana filtro, seguida por la hibridación de una sonda para detectar la secuencia complementaria o casi complementaria). Si se encuentran fragmentos que hibridan en diversas especies relacionadas con la de la sonda (que por lo general se prepara a partir de DNA humano), la sonda se convierte en un candidato para un exón del gen.

Los candidatos son secuenciados; y, si contienen marcos de lectura abiertos, se los utiliza para aislar regiones genómicas de los alrededores. Si éstas parecen ser parte de un exón, pueden ser utilizadas para identificar al gen entero, para aislar el cDNA correspondiente (el DNA obtenido por transcripción inversa a partir del mRNA) o el mRNA mismo y, en última instancia, para identificar la proteína.

Cuando una enfermedad humana es causada por un cambio en una proteína conocida, el gen responsable puede ser identificado debido a que codifica para la proteína, y su responsabilidad sobre la enfermedad puede ser confirmada al demostrar que presenta mutaciones en el DNA de pacientes, pero no en el DNA de individuos no afectados. Sin embargo, en muchos casos, no se conoce la causa de una enfermedad a nivel molecular, y es necesario identificar el gen sin información acerca de su producto proteico.

El criterio básico para la identificación de un gen que está involucrado en una enfermedad humana es demostrar que, en cada paciente con la enfermedad, el gen tiene una mutación que no está presente en el DNA normal. Sin embargo, el gran polimorfismo presente entre los genomas individuales implica que se puedan encontrar muchos cambios cuando se compara el DNA de un paciente con el DNA normal. Antes de la secuenciación del genoma humano, el ligamiento genético podía utilizarse para identificar una región que contuviera un gen de enfermedad, pero la región podría contener muchos genes candidatos. En un gen muy grande, con intrones diseminados en una distancia larga del genoma, era difícil identificar las mutaciones críticas en los pacientes. La disponibilidad de los mapas de SNP de alta resolución y de la secuencia genómica facilitó, en gran medida, la localización precisa de una región más pequeña que contuviera el gen en la cual se pueda comparar, directamente, las secuencias del DNA normal con el del paciente.

Un ejemplo del proceso mediante el cual un gen de enfermedad puede ser rastreado fue proporcionado por el gen responsable para la *distrofia muscular de Duchenne* (DMD), un desorden degenerativo del músculo que está ligado al cromosoma X y que afecta a 1 de cada 3.500 hombres. Los pasos para identificar al gen se resumen en la **FIGURA 4.9**.

El análisis de ligamiento localizó el locus de DMD en la banda cromosómica Xp21. Los pacientes con la enfermedad, a menudo, tienen reordenamientos cromosómicos que involucran a esta banda. Mediante la comparación de la capacidad de las sondas de DNA ligadas al X, de hibridar con el DNA de pacientes y el DNA normal, se obtuvieron los fragmentos clonados que corresponden a la región reordenada o eliminada en el DNA de los pacientes.

Una vez que se ha obtenido el DNA en la vecindad general del gen diana, fue posible realizar el “paseo” a lo largo del cromosoma hasta alcanzar el gen. Se utilizó un **paseo cromosómico** para construir un mapa de restricción de la región a cada lado de la sonda, que cubría una región de > 100 kb. El análisis del DNA, a partir de una serie de pacientes, identificó deleciones grandes en esta región que se extendían en cualquier dirección. La deleción más significativa es una que está contenida enteramente dentro de esta región, debido a que la deleción delinea un segmento que debe ser importante en la función del gen y que indica que el gen –o al menos una parte de él– se encuentra en esa región.

► transferencia zoo (zoo blot)

Uso de la transferencia Southern para evaluar la capacidad de una sonda de DNA de una especie, para formar híbridos con el DNA de los genomas de varias especies diferentes.

► paseo cromosómico

Técnica para ubicar un gen que utiliza los marcadores ligados más cercanos como sonda en una genoteca.

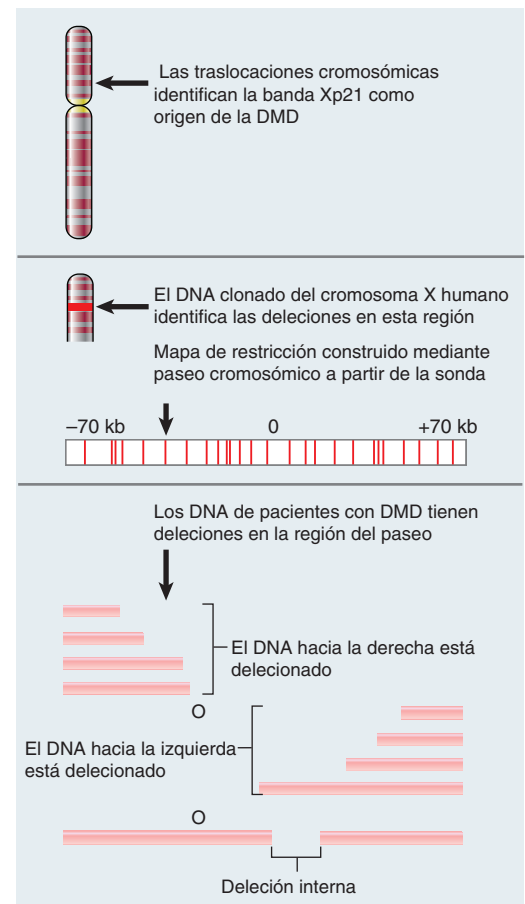


FIGURA 4.9 El gen involucrado en la distrofia muscular de Duchenne fue rastreado mediante mapeo y “paseo cromosómico” en una región en la que se pueden identificar deleciones relacionadas con la ocurrencia de la enfermedad.

Después de identificar la región del gen, fue necesario identificar sus exones e intrones. Mediante la transferencia zoo (zoo blot), se identificaron los fragmentos que tienen hibridación cruzada con el cromosoma X del ratón y con otros DNA de mamífero. Como se resume en la **FIGURA 4.10**, éstos fueron analizados para identificar los marcos de lectura abiertos y las secuencias típicas de las uniones exón-intrón. Los fragmentos que cumplían con estos criterios se utilizaron como sondas para identificar secuencias homólogas en una biblioteca de cDNA preparada a partir de mRNA de músculo.

El cDNA correspondiente al gen identifica un mRNA inusualmente largo (14 kb). La hibridación hacia atrás en el genoma demuestra que el mRNA está codificado por > 60 exones, que se encuentran diseminados en ~2.000 kb de DNA. Esto hace que el gen DMD sea el gen más largo que se haya identificado.

El gen codifica una proteína de ~500 kD denominada *distrofina*, que es un componente del músculo que está presente en cantidades bastante bajas. Todos los pacientes con distrofia muscular de *Duchenne* tienen delecciones en este locus y carecen de distrofina (o ésta es defectuosa).

El músculo también tiene la distinción de tener la proteína más grande, titina, con casi 27.000 aminoácidos. El gen *titina* tiene la mayor cantidad de exones (178) y el exón más largo (17.000 pb) del genoma humano.

Otra técnica que permite analizar rápidamente la presencia de exones en los fragmentos genómicos se denomina **atrapamiento de exones**. La **FIGURA 4.11** muestra que esta técnica comienza con un vector que contiene un promotor fuerte y tiene un único intrón entre dos exones. Cuando se transfecta este vector adentro de células, su transcripción genera cantidades grandes de un RNA que contiene las secuencias de los dos exones. Dentro del intrón, se encuentra un sitio de restricción, el cual es utilizado para insertar fragmentos genómicos de una región de interés. Si un fragmento no contiene un exón, no existirá cambio en el patrón de corte y empalme, y el RNA contendrá las mismas secuencias que el vector parental. Sin embargo, si el fragmento genómico contiene un exón

► atrapamiento de exones

Insertión de un fragmento genómico en un vector, cuya función depende de la provisión de uniones de corte y empalme por parte del fragmento.

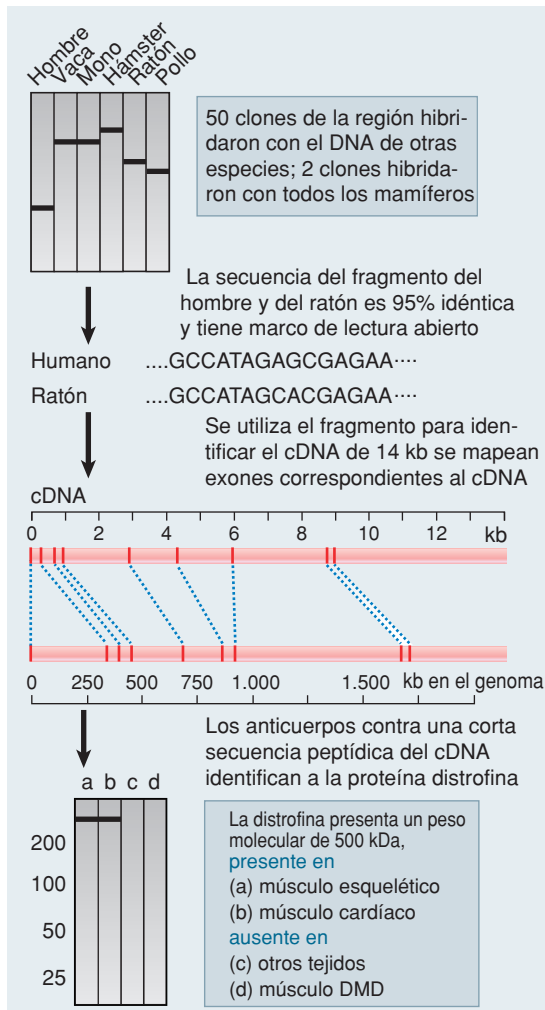


FIGURA 4.10 El gen de la distrofia muscular de *Duchenne* se caracterizó mediante las técnicas de zoo blot, hibridación de cDNA, hibridación genómica y por identificación de la proteína.

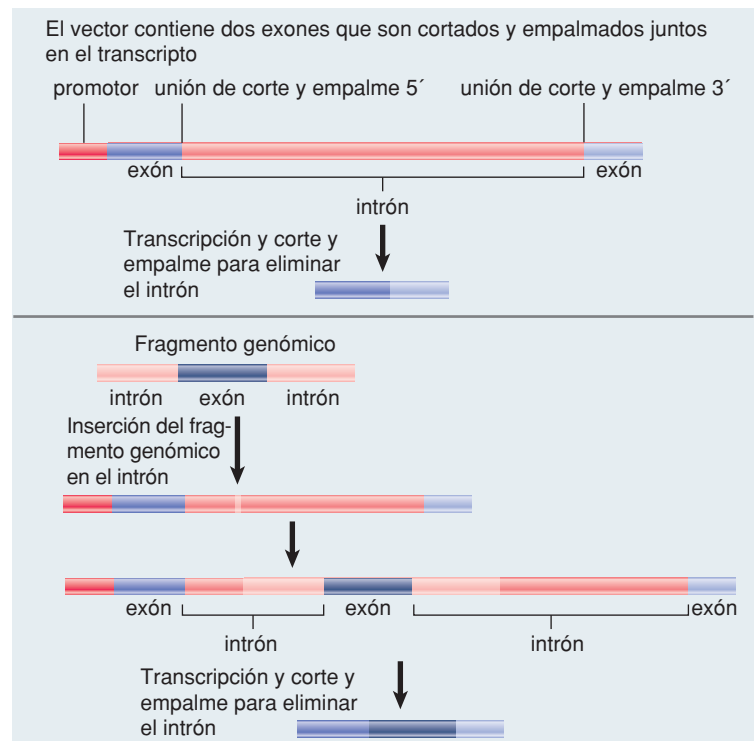


FIGURA 4.11 Para el atrapamiento de exones, se utiliza un vector especial de corte y empalme. Si está presente un exón en el fragmento genómico, su secuencia será recuperada en el RNA citoplasmático. Si el fragmento genómico está compuesto solamente por secuencias del interior del intrón, no se producirá, por lo tanto, el corte y empalme ni se exportará el mRNA al citoplasma.

flanqueado por dos secuencias intrónicas parciales, los sitios de corte y empalme a ambos lados de este exón serán reconocidos y la secuencia del exón se insertará dentro del RNA entre los dos exones del vector. Esto puede detectarse fácilmente mediante la transcripción inversa del RNA citoplasmático a cDNA y utilizando PCR (denominada *RT-PCR*, véase la siguiente sección) para amplificar las secuencias entre los dos exones del vector. De esta manera, se observará que la aparición de las secuencias del fragmento genómico en la población amplificada indican que un exón ha sido “atrapado”. En los genes de mamífero que codifican proteínas, los intrones suelen ser grandes y los exones pequeños; por lo tanto, existe una alta probabilidad de que un fragmento aleatorio del DNA genómico contenga la estructura requerida de un exón rodeado por intrones parciales. De hecho, el atrapamiento de exones puede imitar los acontecimientos que ocurrieron naturalmente durante la evolución de los genes (véase la Sección 3.7, *¿Cómo evolucionaron los genes interrumpidos?*).

CONCEPTOS CLAVE

- La conservación de exones puede utilizarse como fundamento para identificar las regiones codificantes mediante el reconocimiento de fragmentos cuyas secuencias están presentes en múltiples organismos.
- Los genes de enfermedades humanas se identifican mediante el mapeo y la secuenciación del DNA de pacientes para encontrar diferencias respecto del DNA normal que estén genéticamente ligadas a la enfermedad.

REVISIÓN DE CONCEPTOS

La sonda de un exón particular puede identificar múltiples sitios en el genoma. ¿Por qué?

4.8 La conservación de la organización del genoma contribuye a identificar genes

Una vez que se ha determinado la secuencia de un genoma, aún se tiene que identificar los genes dentro de él. Las secuencias codificantes representan una fracción muy pequeña del genoma total. Los potenciales exones pueden identificarse como marcos de lectura abiertos sin interrupciones flanqueados por secuencias adecuadas. ¿Qué criterio se necesita cumplir para identificar un gen funcional (intacto) a partir de una serie de exones?

La **FIGURA 4.12** muestra que un gen funcional debería consistir en una serie de exones para los cuales el primer exón se encontraría inmediatamente a continuación de un promotor; los exones internos, flanqueados por uniones de corte y empalme adecuadas; el último exón, seguido por señales de procesamiento 3'; y se podría deducir un único marco de lectura abierto, el cual comienza con un codón de iniciación y que finaliza con un codón de terminación mediante la unión de todos los exones. En los casos más simples, el primer y el último exón contienen el principio y el final de la región codificante, respectivamente (así como las regiones 3' y 5' no traducidas). En los casos más complejos, el primer o el último exón pueden tener sólo regiones no traducidas y, por lo tanto, pueden ser más difíciles de identificar.

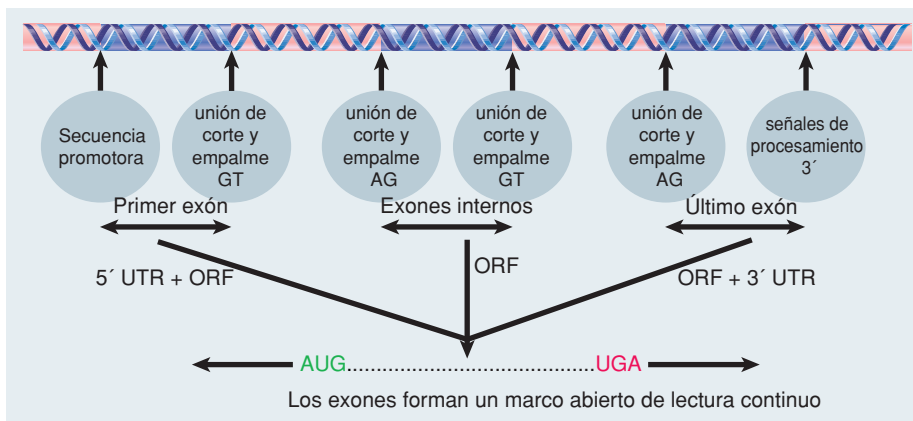


FIGURA 4.12 Los exones de genes que codifican proteína se identifican como secuencias codificante flanqueadas por las señales adecuadas con regiones no traducidas en ambos extremos. La serie de exones debe generar un marco de lectura abierto con los codones de iniciación y de terminación apropiados.

Los algoritmos que se utilizan para conectar los exones no son completamente eficientes cuando el genoma es muy grande y los exones están separados por distancias muy grandes.

Por ejemplo, en el análisis inicial del mapa del genoma humano se contabilizaron 170.000 exones dentro de 32.000 genes. Esto probablemente sea incorrecto debido a que daría un promedio de 5,3 exones por gen, mientras que el promedio de exones en los genes individuales que han sido completamente caracterizados es de 10,2. O se han omitido exones, o ellos deberían estar conectados de manera diferente en una cantidad más pequeña de genes en la secuencia del genoma completo.

Incluso cuando la organización de un gen se identifica de manera correcta, existe el problema de distinguir los genes funcionales de los pseudogenes. Muchos pseudogenes pueden reconocerse por defectos obvios dado que presentan mutaciones múltiples, que crean una secuencia codificante no funcional. Los pseudogenes que surgieron más recientemente no han acumulado tantas mutaciones y, por lo tanto, puede ser más difícil reconocerlos. En un ejemplo extremo, el ratón tiene sólo un gen *Gapdh* funcional (que codifica para la gliceraldehído fosfato deshidrogenasa), pero tiene ~400 pseudogenes. Aproximadamente 100 de éstos inicialmente parecían ser funcionales en la secuencia del genoma del ratón, y fue necesario un análisis individual para excluirlos de la lista de genes funcionales.

¿Cómo se puede verificar un gen putativo que codifica una proteína? Si se puede demostrar que una secuencia de DNA se transcribe y procesa a un mRNA traducible, se puede asumir que es funcional. Una técnica para realizar esta verificación es la **transcripción inversa-reacción en cadena de la polimerasa (RT-PCR)**, en la que el RNA aislado –a partir de células– es transcrito de forma inversa a DNA y, posteriormente, es amplificado a muchas copias mediante la utilización de la reacción en cadena de la polimerasa. Los productos amplificados del DNA después pueden ser secuenciados o analizados de otra forma para comprobar si tienen las características estructurales apropiadas de un transcripto maduro. La RT-PCR también puede utilizarse como una prueba cuantitativa de la expresión génica.

La confianza de que un gen es funcional puede incrementarse al comparar las regiones de los genomas de diferentes especies. Ha existido una gran reorganización global de las secuencias entre los genomas del humano y del ratón, como se puede ver en el simple hecho de que existen 23 cromosomas en el genoma haploide humano y 20 cromosomas en el genoma haploide del ratón. Sin embargo, a nivel local el orden de los genes en general es el mismo: cuando se comparan los pares homólogos humano y ratón, los genes localizados en ambos lados tienden a ser homólogos. Esta relación se denomina **sintenia**.

La **FIGURA 4.13** muestra la relación entre el cromosoma 1 de ratón y el conjunto de cromosomas humanos. Se pueden reconocer 21 segmentos en este cromosoma de ratón, que tienen contrapartes sinténicas en los cromosomas humanos. El grado del barajado de exones que ha ocurrido entre los genomas está demostrado por el hecho de que los segmentos se encuentran diseminados entre seis cromosomas humanos diferentes. Los mismos tipos de relaciones se encuentran en todos los cromosomas de ratón, excepto para el cromosoma X, que es sinténico sólo con el cromosoma X humano. Esto se explica por el hecho de que el X es un caso especial, y sujeto a una compensación de la dosis para que se ajuste la diferencia entre una copia de los machos y las dos copias de las hembras (véase la *Sección 27.5, El cromosoma X experimenta cambios globales*). Esta restricción puede aplicar presión selectiva contra la translocación de los genes hacia el cromosoma X y desde él.

La comparación de las secuencias del genoma humano y del ratón muestran que > 90% de cada genoma se encuentra en bloques de sintenia que varían ampliamente de tamaño, desde 300 kb hasta 65 Mb. Existe un total de 342 segmentos sinténicos, con una longitud promedio de 7 Mb (0,3% del genoma). El 99% de los genes del ratón tiene un homólogo en el genoma humano; para el 96% ese homólogo está en una región de sintenia.

La comparación de los genomas proporciona información interesante acerca de la evolución de las especies. El número de familias de genes en el genoma humano y en el del ratón es el mismo, y una diferencia importante entre las especies es la expansión diferencial de las familias particulares en el genoma del ratón. Esto es especialmente notable en los genes que afectan características fenotípicas que son únicas a las especies. De las 25 familias para las cuales el tamaño se ha expandido en el ratón, 14 contienen genes específicamente involucrados en la reproducción del roedor, y 5 contienen genes específicos para el sistema inmunitario.

► **transcripción inversa-reacción en cadena de la polimerasa (RT-PCR)** Una técnica para la detección y cuantificación de la expresión de un gen mediante la transcripción inversa y la amplificación del RNA de una muestra celular.

► **sintenia** Relación entre regiones cromosómicas de distintas especies, en las que aparecen genes homólogos en el mismo orden.

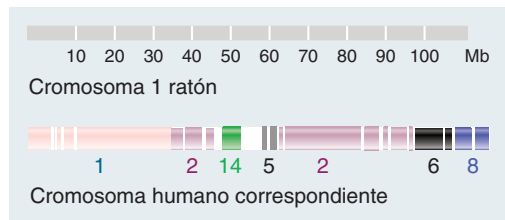


FIGURA 4.13 El cromosoma 1 de ratón tiene 21 segmentos –de 1 a 25 Mb– que son sinténicos con regiones que corresponden a las partes de seis cromosomas humanos.

Una comprobación de la importancia de la identificación de los bloques de sintenia está dada por la comparación de a pares de los genes que se encuentran dentro de ellos. Por ejemplo, un gen que no está en una localización sinténica (es decir, su contexto es diferente en las dos especies que se comparan) tiene el doble de probabilidad de ser un pseudogén. De otra forma, la translocación fuera del locus original tiende a estar asociada con la formación de pseudogenes. La falta de un gen relacionado en una posición de sintenia es, por lo tanto, la base para sospechar que un gen aparente puede realmente ser un pseudogén. En términos generales, > 10% de los genes que inicialmente se identifican por análisis del genoma es probable que resulten ser pseudogenes.

Como regla general, las comparaciones entre los genomas suman efectividad a la predicción génica de manera significativa. Cuando las características de una secuencia indican que los genes activos se conservan, como por ejemplo entre los genomas humano y de ratón, existe una mayor probabilidad de que éstos identifiquen ortólogos activos.

La identificación de genes que codifican RNA, además de los mRNA, es más difícil debido a que no se puede utilizar el criterio del marco de lectura abierto. Es cierto que el análisis comparativo del genoma descrito anteriormente ha incrementado el rigor del análisis. Por ejemplo, el análisis separado de genoma humano o del ratón identifica ~500 genes que codifican tRNA, pero las comparaciones de las características sugieren que < 350 de estos genes son en realidad funcionales en cada genoma.

Un gen activo puede localizarse mediante el uso de una **etiqueta de secuencia expresada (EST)**—una porción corta de una secuencia transcrita, en general obtenida a partir de la secuenciación de uno o más extremos de un fragmento clonado a partir de una biblioteca de cDNA. Una etiqueta de secuencia expresada puede confirmar que un gen candidato se transcribe realmente o que puede ayudar a identificar genes que influyen en desordenes particulares. A través del uso de una técnica de mapeo físico, como la hibridación *in situ* (véase la Sección 6.9, *Los DNA satélites suelen encontrarse en la heterocromatina*), se puede determinar la localización cromosómica de una etiqueta de secuencia expresada.

► **etiqueta de secuencia expresada (EST)** Un fragmento corto secuenciado de un cDNA que puede utilizarse para identificar un gen que se expresa activamente.

CONCEPTOS CLAVE

- Los métodos para identificar los genes activos no son perfectos y se deben hacer muchas correcciones hasta llegar a las estimaciones preliminares.
- Se debe distinguir a los pseudogenes de los genes activos.
- Existen relaciones sinténicas entre los genomas humano y del ratón, y la mayoría de los genes activos se encuentran en una región sinténica.

REVISIÓN DE CONCEPTOS

En ausencia de información definitiva acerca de la expresión de una secuencia, como la presencia de un producto génico funcional, ¿qué características de una secuencia eucarionte podrían sugerir fuertemente qué es un gen activo que codifica un polipéptido?

4.9 Algunos orgánulos tienen DNA

La primera evidencia de la presencia de genes fuera del núcleo la proporcionó la **herencia no mendeliana** en las plantas (observada en los primeros años del siglo xx, justo después del redescubrimiento de la herencia mendeliana). La herencia no mendeliana se define por el fracaso de la progenie de un apareamiento para exhibir la segregación mendeliana de los caracteres parentales y, por lo tanto, es tomada para indicar la presencia de genes que residen fuera del núcleo y que no utilizan la segregación en los husos mitótico y meiótico para distribuir copias de los gametos o de las células hijas, respectivamente. La **FIGURA 4.14** muestra que esto sucede cuando las mitocondrias heredadas de progenitores masculino y femenino tienen diferentes alelos y una célula hija recibe una distribución desequilibrada de la mitocondria a partir de un solo progenitor (véase la Sección 17.11, *¿Cómo se replica y segregan las mitocondrias?*). Esto también es un hecho verdadero para los cloroplastos de las plantas, ya que tanto la mitocondria como el cloroplasto contienen genomas con genes funcionales (véase a continuación).

La forma extrema de una herencia no mendeliana es la herencia uniparental, que ocurre cuando se hereda el genotipo de un solo progenitor y el del otro progenitor no se tras-

► **herencia no mendeliana** Un patrón de herencia que no sigue el esquema esperado según los principios de Medel (cada progenitor contribuye con un solo alelo a la descendencia). Los genes extranucleares muestran un patrón de herencia no mendeliana.

MÉTODOS Y TÉCNICAS

El uso del mtDNA para reconstruir filogenias humanas

Desde que a comienzos de la década de 1960 se descubrió que las mitocondrias contienen DNA, este mtDNA ha llamado mucho la atención. En los animales, el mtDNA es relativamente pequeño, sólo alrededor de 16,6 kb, por lo que lo convierte en un genoma simple e ideal para analizar. Este genoma pequeño codifica diversos genes que son absolutamente esenciales para la función mitocondrial, pero no contiene intrones y tiene muy poca secuencia intergénica. La única región que tiene variabilidad significativa es el bucle D, o región control. El bucle D está involucrado en la regulación de la transcripción y la replicación del DNA y constituye cerca de 1.122 pb del genoma (con algunas variaciones en la longitud). La región control no codifica genes, es altamente variable y evoluciona rápidamente en relación con el resto del genoma mitocondrial. El mtDNA humano, y el mtDNA animal en general, tiene una tasa de mutación mucho más alta que la tasa promedio de los genes nucleares, mientras que la organización (el orden de los genes en el genoma) es altamente conservada. El mtDNA es, por lo tanto, un buen blanco para identificar diferencias individuales aun dentro de las poblaciones.

Dado el tamaño pequeño y la naturaleza altamente conservada del mtDNA animal, es relativamente simple secuenciar ya sea la molécula entera o porciones de ella, y el análisis del mtDNA ha sido un elemento importante de los estudios filogenéticos humanos durante casi tres décadas [del siglo xx]. El mtDNA es un objetivo lógico que simplifica el análisis genético debido a la herencia uniparental y a la falta relativa de recombinación entre las variantes de este genoma de orgánulo. En contraste, el análisis de los genes autosómicos carece de la claridad de los marcadores uniparentales (el genoma mitocondrial, a partir de la madre, y el cromosoma Y, a partir del padre).

El primer análisis de la variación del mtDNA utilizó la digestión con enzimas de restricción y caracterización de la molécula entera de mtDNA (publicado en 1981 por Anderson y cols.). Este estudio informó la capacidad del análisis de restricción para trazar la historia genética humana a partir de un grupo de 21 personas de diversas etnias y regiones geográficas, sobre la base de la observación del *polimorfismo de longitud en los fragmentos de restricción* (RFLP). Esto condujo a la estimación de alrededor de 180.000 años para la variación global del mtDNA —el lapso de tiempo transcurrido para la aparición de todas las variaciones a partir de un punto inicial en común—, que es notablemente consistente con respecto a los estudios más recientes.

Hacia la década de los noventa [del siglo xx], el análisis de los RFLP se complementó con el análisis de la secuencia de los *segmentos hipervariables* (SHV) de la región control del mtDNA para proporcionar información adicional para

el análisis filogenético. La porción de SHV-I (cerca de 350 pb) de la región control es altamente variable, pero es susceptible a mutaciones recurrentes que pueden tornar dificultoso el análisis. Las regiones pequeñas, tales como éstas, pueden ser amplificadas mediante PCR y secuenciadas de rutina para obtener información para la construcción de árboles filogenéticos. En la práctica, sólo unas pocas regiones del mtDNA son blanco para la amplificación y análisis de secuencia, principalmente dentro de los segmentos SHV. Se ha estimado que la tasa de divergencia de los segmentos hipervariables de la región control del mtDNA está entre el 11,5 y el 17,3% por un millón de años. Mediante el análisis de secuencias que se suman al análisis por RFLP, se ha producido un árbol filogenético humano bastante minucioso, con las ramificaciones principales para las diferentes regiones geográficas del mundo. Cuanta más información se disponga, más minucioso se tornará el árbol.

Como resultado del análisis de la región control del mtDNA, se han caracterizado muchos haplogrupos; es decir, individuos de diferentes partes del mundo tienen tipos de mtDNA distintivos y, por ende, se los coloca en diferentes haplogrupos. Luego se asocia a cada haplogrupo con una región particular del mundo. Los principales haplogrupos identificados son, en términos generales, África sub-sahariana, Asia del Este, Asia del Sur, Oceanía, Europa y América. Las personas que migran de una región a otra llevan su tipo de mtDNA con ellos, por lo que proporcionan una etiqueta para el análisis geográfico y genealógico. Esto proporcionó las bases para determinar dónde vivió el último ancestro humano común de los tipos de mtDNA modernos.

El análisis de mtDNA condujo a la generación de un árbol filogenético relativamente complejo, con orígenes en África hace aproximadamente 200.000 años. Puesto que el genoma mitocondrial es de herencia materna, algunos han referido como “la Eva mitocondrial” a la raíz o al ancestro común más reciente de los haplotipos del mtDNA. Con el transcurso del tiempo, han descendido muchos haplogrupos, con muchos nodos que divergieron hace aproximadamente de 80.000 a 90.000 años, y que pueden corresponder a cambios en el clima después de una fase interestadial glacial. Otros nodos pueden estar asociados con la migración y la colonización de varias zonas del mundo. Este enfoque se ha utilizado para determinar que los nativos de América se originaron en Asia. En otros estudios, se ha demostrado que los nativos europeos comparten esencialmente los mismos haplogrupos con las personas del cercano oriente, pero no con los africanos o los asiáticos del Este. Muchos, pero no todos, de los descubrimientos muestran una fuerte correlación con los análisis de marcadores genéticos en el cromosoma Y, transmitido a través del padre.

A medida que mejora la tecnología para secuenciar y que disminuye el costo para obtener información de secuencias de DNA, se ha tornado más factible secuenciar la mo-

lécua enteras de mtDNA, en lugar de sólo segmentos pequeños. Esto ha permitido un análisis más completo y preciso, que simplifica el análisis genético, del cual resulta un árbol de mtDNA con una resolución filogeográfica mucho más alta, a la vez que proporciona un mayor refinamiento de los orígenes geográficos. Por ejemplo, el análisis de la secuencia completa del mtDNA condujo a la conclusión de que los agrupamientos de haplogrupos de Asia Central y de los nativos de América sean subconjuntos de la variación del mtDNA de Asia del Este. Esta distinción no sería posible sólo a través de la comparación de las secuencias de SHV. Sin embargo, el análisis de la secuencia completa del mtDNA genera cierto desafío cuando se compara la información de secuencia de tramos cortos de mtDNA –que se obtienen a partir de muestras de DNA ancestral de huesos, dientes o de otras muestras–, con la información del genoma mitocondrial completo de muestras actuales. Las muestras de DNA ancestral con frecuencia carecen de la suficiente integridad que permita el análisis de una secuencia completa del mtDNA.

Las interpretaciones hechas sobre la base de la información de la secuencia del mtDNA podrían verse afectadas y generar incertidumbre acerca de si el mtDNA sufre (o no) recombinación homóloga. La recombinación del mtDNA se produce fácilmente en levaduras, hongos y plantas. Sin embargo, se cree que en los animales no se produce recombinación del mtDNA –al menos parcialmente, debido al tamaño extremadamente pequeño del genoma y a la herencia uniparental de la madre. La ausencia de recombinación simplifica el análisis al permitir la máxima resolución molecular simplemente como una función de la longitud de secuencia examinada, en lugar de la necesidad de tomar en cuenta la variación de la estructura génica como resultado de la recombinación, que es común para los genes autosómicos nucleares. Muchos investigadores han informado que se puede producir la recombinación de mtDNA a niveles bajos y en algunos órganos específicos de seres humanos y de otros animales. Sin embargo, esto de alguna manera es controvertido y requiere experimentos posteriores que determinen el grado de la recombinación en el mtDNA animal. Si se produce la recombinación del mtDNA, entonces ésta debe ser considerada, y puede requerirse cierta alteración de los árboles filogenéticos.

Además del análisis filogenético, la información de la secuencia de mtDNA ha sido útil para caracterizar las enfermedades mitocondriales espontáneas y hereditarias. Durante las últimas dos décadas [del siglo xx] se han caracterizado una cantidad de enfermedades del mtDNA, que incluyen las miopatías mitocondriales que conducen al deterioro muscular, y la neuropatía óptica hereditaria, que conduce a la ceguera. Estas enfermedades se deben a mutaciones en los genes codificados en el mtDNA. Puesto que el genoma mitocondrial humano es tan compacto, y con muy poca secuencia no codificante, casi cualquier mutación

puede llevar a un defecto si la mutación causa un cambio en un aminoácido que está incorporado en la proteína correspondiente. Muchas de estas enfermedades son el resultado de *deleciones* de la secuencia codificante. De hecho, ésta es la razón de por qué las deleciones suelen generar “anticipación genética”; la variante defectiva surge como una copia entre muchas, pero la frecuencia dentro de una célula puede incrementarse gradualmente debido a que es más corta y se replica más rápido. Como resultado, la enfermedad puede tornarse cada vez más grave en un individuo en el transcurso del tiempo (como se mencionó); y, si es en las células germinales, puede presentarse más temprano en las siguientes generaciones. Sin embargo, puesto que en una célula existen múltiples mitocondrias, los defectos se tornan aparentes sólo cuando las moléculas con una mutación particular están multiplicadas y se acumulan. Por lo tanto, muchas de estas enfermedades son progresivas, en lugar de tener un inicio repentino. Se han asociado muchos otros desórdenes con defectos en el mtDNA, como en la encefalomiopatía mitocondrial, la distonía, algunos casos de diabetes y de la enfermedad de Alzheimer y también el envejecimiento. Algunas de estas enfermedades son hereditarias, mientras que otras parecen estar causadas por factores ambientales. Para las enfermedades mitocondriales hereditarias, las mismas ventajas de trabajar con moléculas pequeñas con poco o nada de recombinación permiten el análisis de patrones de herencia.

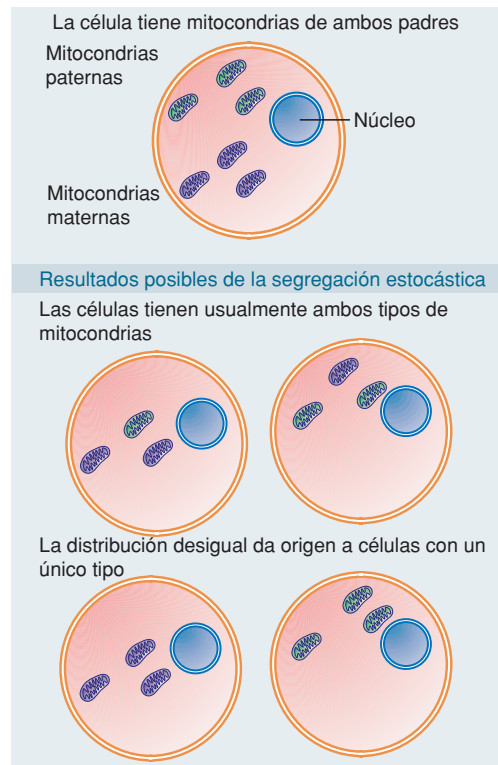
Aunque el análisis de secuencia del mtDNA es muy exitoso para el estudio de las filogenias de mtDNA humano, no es adecuado para el análisis de mtDNA de plantas y hongos. Los genomas mitocondriales en las plantas y hongos son más grandes, todavía existen regiones control no identificadas –dada la gran diferencia en la organización génica–, y son comunes los intrones y espacios intergénicos largos. Además, el mtDNA de plantas ha demostrado tener una tasa de mutación mucho menor, en comparación con la del mtDNA animal, pero una tasa de reordenamiento mucho más alta. Aún es posible, y de hecho es bastante común, el análisis de secuencia de genes o de regiones no codificantes de mtDNA de estos organismos para llevar a cabo estudios filogenéticos.

Fuentes y lecturas recomendadas

- Kraytsberg et al. (2004). Recombination of human mitochondrial DNA. *Science* 304:981.
- Torroni et al. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* 22:339-345.
- Underhill and Kivisild (2007). Use of chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics* 41:539-564.
- Vigilant et al. (1991). African populations and the evolution of human mitochondrial DNA. *Science* 253:1503-1507.
- Wallace (1997). Mitochondrial DNA in aging and disease. *Scientific American*, August:40-47.

FIGURA 4.14 Cuando los alelos mitocondriales materno y paterno difieren, se debe a que una célula tiene dos conjuntos de DNA mitocondrial. La mitosis suele generar células hijas con ambos conjuntos. La variación somática puede surgir si la segregación desigual genera células hijas sólo con un conjunto.

► **herencia materna** La sobrevivencia preferencial en la progenie de los marcadores genéticos aportados por uno de los padres.



mite a la descendencia. En los ejemplos más comunes, un genotipo progenitor excede al otro genotipo en la progenie. En los animales y en la mayoría de las plantas, se hereda de manera preferencial (o únicamente) el genotipo de la madre. Este efecto algunas veces se describe como **herencia materna**. La cuestión importante es que predomina el genotipo contribuido por el progenitor de un sexo particular, como se observa en las proporciones de segregación cuando se realiza una cruce entre mutante y tipo silvestre. Esto contrasta con el comportamiento de la genética mendeliana, que se produce cuando las cruces recíprocas muestran las contribuciones de ambos progenitores que han de ser igualmente heredadas.

El sesgo en los genotipos parentales se establece en el cigoto, o justo después de su formación. Existen varias causas posibles. La contribución de la información materna o paterna a los orgánulos del cigoto puede ser desigual; en el caso más excepcional, sólo contribuye un progenitor. En otros casos, las contribuciones son equitativas, pero

la información proporcionada por un progenitor no sobrevive. Las combinaciones de ambos efectos son posibles. Cualesquiera sean las causas, la representación desigual de la información a partir de dos progenitores contrasta con la información genética nuclear, que deriva por igual de cada progenitor.

Parte de la herencia no mendeliana es el resultado de la presencia, en las mitocondrias o en los cloroplastos, de los genomas de DNA que se heredan independientemente de los genes nucleares. En efecto, el genoma del orgánulo abarca una longitud del DNA que ha sido físicamente secuestrado en una parte definida de la célula y que está sujeto a su propia forma de expresión y regulación. Un genoma de orgánulo puede codificar algunos o todos los tRNA y rRNA, pero codifica sólo algunos polipéptidos necesarios para perpetuar el orgánulo. Los otros polipéptidos están codificados en el núcleo, expresados a través del aparato de síntesis de proteína citoplasmático, e importado hacia el interior del orgánulo.

► **genes extranucleares** Genes que residen fuera del núcleo; en orgánulos como la mitocondria y el cloroplasto.

Los genes que no residen dentro del núcleo suelen describirse como **genes extranucleares**; ellos se transcriben y traducen en el mismo orgánulo (mitocondria o cloroplasto) en el cual residen. Por el contrario, los genes nucleares se expresan por medio de la síntesis proteica citoplasmática. (El término “herencia citoplasmática” suele utilizarse para describir el comportamiento de los genes en los orgánulos. No obstante, no se utilizará esta descripción, debido a que es importante poder distinguir entre los acontecimientos que ocurren en el citosol de aquellos que ocurren en los orgánulos específicos.)

Los animales muestran herencia materna de la mitocondria: esto puede explicarse cuando es el óvulo el que aporta completamente las mitocondrias, mientras que el espermatozoide no contribuye para nada. En la **FIGURA 4.15** se muestra que el espermatozoide contribuye sólo con una copia del DNA nuclear. Así, los genes mitocondriales derivan exclusivamente de la madre y en los machos, se descartan en cada generación. Los cloroplastos en general son de herencia materna, aunque algunos taxones de plantas muestran herencia paterna o biparental de los cloroplastos.

El ambiente químico de los orgánulos es diferente al del núcleo y por lo tanto, el DNA del orgánulo tiene su propia tasa de evolución. Si la herencia es uniparental, no puede existir recombinación entre los genomas parentales. De hecho, la recombinación no suele ocurrir en aquellos casos para los cuales los genomas de los orgánulos son heredados de ambos progenitores. El DNA del orgánulo tiene un sistema de replicación diferente al del

núcleo; como resultado, la tasa de error durante la replicación puede ser diferente. En los mamíferos, el DNA mitocondrial acumula mutaciones más rápidamente que el DNA nuclear; pero, en las plantas, la acumulación de mutaciones es más lenta en la mitocondria que en el núcleo: mientras que el DNA del cloroplasto tiene una tasa de mutación intermedia.

Una consecuencia de la herencia materna es que la secuencia del DNA mitocondrial es más sensible que la del DNA nuclear a las reducciones en el tamaño de la población reproductiva. Las comparaciones de las secuencias de DNA mitocondrial en un espectro de poblaciones humanas permiten construir un árbol evolutivo. La divergencia entre los DNA mitocondriales humanos abarca el 0,57%. Se puede construir un árbol en el cual las variantes mitocondriales divergieron a partir de un ancestro común (africano). La tasa a la cual el DNA mitocondrial de los mamíferos acumula mutaciones es del 2 a 4% por millón de años, que es > 10 veces más rápido que la tasa para la globina. Una tasa como ésta podría generar la divergencia observada en un período evolutivo de 140.000 a 280.000 años. Esto implica que el DNA mitocondrial humano descende de una población única que vivió en África hace ~200.000 años. Sin embargo, esto no puede interpretarse como evidencia de que existió una única población al mismo tiempo, ya que pudieron existir muchas poblaciones, y algunas de ellas, o todas, pudieron haber contribuido con la variación genética nuclear humana moderna.

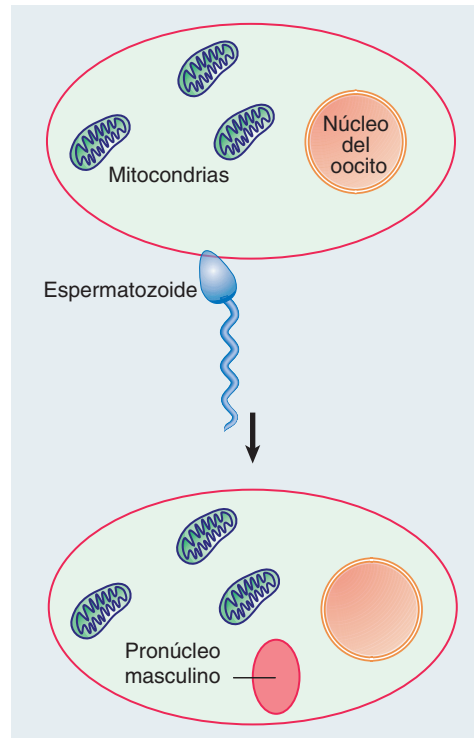


FIGURA 4.15 El DNA del espermatozoide entra en el oocito para formar el pronúcleo masculino en el óvulo fecundado, pero todas las mitocondrias son proporcionadas por el oocito.

CONCEPTOS CLAVE

- Las mitocondrias y los cloroplastos tienen genomas que exhiben herencia no mendeliana. En general, son de herencia materna.
- Los genomas de los orgánulos pueden sufrir segregación somática en las plantas.
- Las comparaciones del DNA mitocondrial sugieren que descende de una sola población que vivió hace 200.000 años en África.

REVISIÓN DE CONCEPTOS

¿Por qué la tasa de mutación del DNA mitocondrial en los mamíferos puede ser un orden de magnitud mayor que la del DNA nuclear? (Indicio: ¿qué compuestos se encuentran en las mitocondrias y no en el núcleo?)

4.10 Los genomas de los orgánulos son DNA circulares que codifican las proteínas de los orgánulos

La mayoría de los genomas de los orgánulos toma la forma de una molécula circular de DNA de secuencia única (denotada **mtDNA** en la mitocondria y **ctDNA** o **cpDNA** en el cloroplasto). Existen unas pocas excepciones en los eucariontes unicelulares para las cuales el DNA mitocondrial es una molécula lineal.

Por lo general, existen diversas copias del genoma en el orgánulo individual. También hay múltiples orgánulos por célula; por lo tanto, hay muchos genomas de orgánulos por célula; por lo tanto, el genoma de orgánulo puede considerarse una secuencia repetitiva.

Los genomas de cloroplastos son relativamente grandes, en general ~140 kb en las plantas superiores y < 200 kb en los eucariontes unicelulares. Esto es comparable al tamaño

- ▶ **mtDNA** DNA mitocondrial
- ▶ **ctDNA (cpDNA)** DNA del cloroplasto

del genoma de un bacteriófago grande, como el del T4 de ~165 kb. Existen múltiples copias del genoma por orgánulo –en general, de 20 a 40 en una planta superior–, y múltiples copias del orgánulo por célula, entre 20 y 40.

Los genomas mitocondriales varían en el tamaño total por más de un orden de magnitud. Las células animales tienen genomas mitocondriales pequeños (aproximadamente 16,5 kb en los mamíferos). Existen varios cientos de mitocondrias por célula, y cada mitocondria tiene múltiples copias del DNA. La cantidad total de DNA mitocondrial, en relación con el DNA nuclear, es pequeña; se estima que es < 1%.

En la levadura, el genoma mitocondrial es mucho más grande. En *Saccharomyces cerevisiae*, el tamaño exacto varía entre las diferentes cepas, pero tiene en promedio ~80 kb. Hay ~22 mitocondrias por célula, que corresponden a ~4 genomas por orgánulo. En las células en crecimiento, la proporción de DNA mitocondrial puede ser tan alta como el 18%.

Las plantas exhiben una variación extremadamente amplia en el tamaño del DNA mitocondrial, con un mínimo de ~100 kb. El tamaño del genoma torna dificultoso el aislamiento intacto, pero el mapeo de restricción realizado en diversas plantas sugiere que el genoma mitocondrial suele ser una secuencia única organizada como un círculo. Dentro de este círculo, existen secuencias homólogas cortas. La recombinación entre estos elementos genera moléculas circulares subgenómicas más cortas que coexisten con el genoma “maestro” completo –un buen ejemplo de la aparente complejidad de los DNA mitocondriales de las plantas.

Con los genomas mitocondriales secuenciados a partir de muchos organismos, se pueden apreciar ahora algunos patrones generales en la representación de las funciones en el DNA mitocondrial. La **FIGURA 4.16** resume la distribución de los genes en los genomas mitocondriales. El número total de genes que codifican proteínas es más bien pequeño y no se correlaciona con el tamaño del genoma. Los 16 kb de los genomas mitocondriales de mamífero codifican 13 proteínas, mientras que los 60 a 80 kb de los genomas mitocondriales de levadura codifican apenas 8 proteínas. Los genomas mitocondriales de plantas mucho más grandes codifican más proteínas. En la mayoría de los genomas mitocondriales se encuentran intrones, aunque están ausentes en los genomas muy pequeños de mamífero.

Los dos rRNA principales siempre están codificados por el genoma mitocondrial. La cantidad de tRNA codificada en el genoma mitocondrial varía desde ninguno hasta el complemento completo (25 a 26 en la mitocondria). Esto explica la variación en la cantidad que se observa en la **Figura 4.16**.

La mayor parte de la actividad codificante de proteína corresponde a los componentes de la estructura de multisubunidades de los complejos I-IV de respiración. En los genomas mitocondriales de plantas y protistas, se encuentran codificadas muchas proteínas ribosómicas, pero existen unas pocas o ninguna en los genomas de animales y hongos. En muchos genomas mitocondriales de protistas, existen genes que codifican proteínas involucrados en la importación mitocondrial.

El DNA mitocondrial animal es extremadamente compacto. Existen diferencias de gran magnitud en la detallada organización génica que se encuentran en diferentes taxones animales, pero se mantiene el principio general de un genoma pequeño que codifica para una cantidad restringida de funciones. En las mitocondrias de mamíferos, el genoma es extremadamente compacto. No existen intrones; en realidad, algunos genes se superponen, y casi cada par de bases puede asignarse a un gen. A excepción del **bucle D** –una región involucrada con la iniciación de la replicación del DNA–: no más de 87 de los 16.569 pb del genoma mitocondrial humano se encuentran en regiones intercistrónicas.

Las secuencias nucleotídicas completas de los genomas mitocondriales de los animales exhiben gran homología en su organización. En la **FIGURA 4.17**, se resume el mapa del genoma mitocondrial humano. Existen 13 regiones que codifican proteínas. Todas las proteínas son componentes del sistema de transferencia de electrones de la respiración celular. Éstas incluyen el citocromo b, tres subunidades de la citocromo oxidasa, una de las subunidades de la ATPasa y siete subunidades (o proteínas asociadas) de la NADH deshidrogenasa.

| Especies | Tamaño (kb) | Genes que codifican proteína | Genes que codifican RNA |
|-----------|-------------|------------------------------|-------------------------|
| Hongo | 19–100 | 8–14 | 10–28 |
| Protistas | 6–100 | 3–62 | 2–29 |
| Plantas | 186–366 | 27–34 | 21–30 |
| Animales | 16–17 | 13 | 4–24 |

FIGURA 4.16 Los genomas mitocondriales tienen genes que codifican proteínas (mayormente, complejos I-IV), rRNA y tRNA.

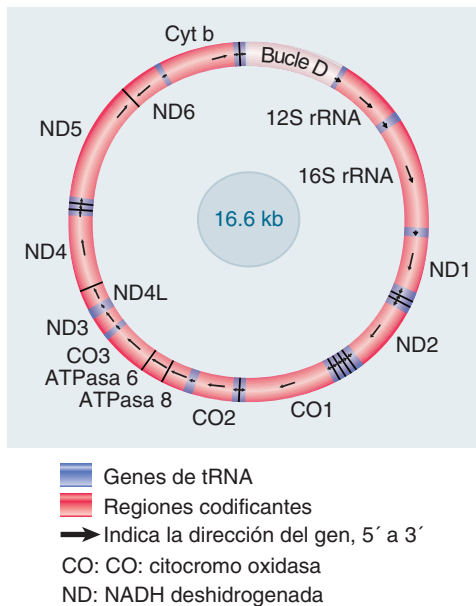


FIGURA 4.17 El DNA mitocondrial humano tiene 22 genes de tRNA, 2 genes de rRNA y 13 regiones codificantes de proteínas. Catorce de las 15 regiones que codifican proteína o rRNA se transcriben en la misma dirección. Catorce de los genes de tRNA se expresan en la dirección del sentido de las agujas de reloj; y 8, en sentido contrario.

► **bucle D** Región del DNA mitocondrial de los animales variable en tamaño y secuencia, que contiene el origen de replicación.

Solo la discrepancia de tamaño en cinco veces entre los genomas mitocondriales de *S. cerevisiae* (84 kb) y de mamíferos (16 kb), alerta acerca de que debe existir una gran diferencia en su organización genética, a pesar de su función común. El número de productos sintetizados de forma endógena relacionados con la función enzimática mitocondrial parece ser similar. ¿Esto significa que el material genético adicional en la mitocondria de la levadura representa a otras proteínas que quizás estén relacionadas con la regulación, o bien, éste no se expresa?

El mapa que se muestra en la **FIGURA 4.18** representa los principales productos proteicos y de RNA de la mitocondria de levadura. La característica más notable es la dispersión del loci en el mapa.

Los dos loci más prominentes son los genes interrumpidos *box* –que codifica para el citocromo b–, y *oxi3* –que codifica para la subunidad 1 de la citocromo oxidasa. ¡Juntos, estos dos genes son casi tan largos como el genoma mitocondrial entero en los mamíferos! Muchos de los intrones largos en estos genes tienen marcos de lectura abiertos en registro con el exón precedente (véase la Sección 29.5, *Algunos intrones del grupo I codifican endonucleasas que promueven la movilidad*). Esto añade diversas proteínas, todas sintetizadas en cantidades bajas, al complemento de la mitocondria de levadura.

Los genes remanentes se encuentran sin interrupciones. Ellos corresponden a las otras dos subunidades de la citocromo oxidasa codificadas por la mitocondria: la de la subunidad(es) de la ATPasa y (en el caso de *var1*) a la proteína ribosómica mitocondrial. El número total de genes mitocondriales de levadura es improbable que excedan los ~25.

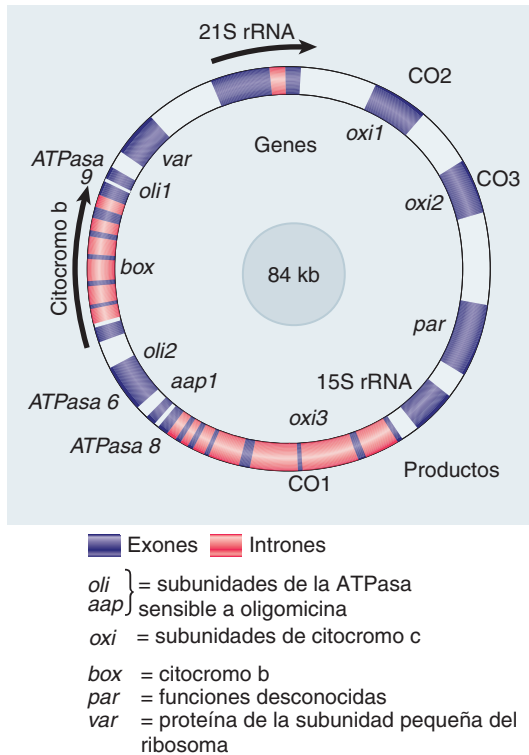


FIGURA 4.18 El genoma mitocondrial de *S. cerevisiae* contiene genes tanto interrumpidos como no interrumpidos, que codifican proteínas, rRNA y tRNA (no se indican las posiciones). Las flechas indican la dirección de la transcripción.

CONCEPTOS CLAVE

- Los genomas de los orgánulos, en general (pero no siempre), son moléculas circulares de DNA.
- Los genomas de orgánulos codifican algunas, pero no todas, las proteínas que se encuentran en los orgánulos.
- El DNA mitocondrial de una célula animal es extremadamente compacto y suele codificar para 13 proteínas, 2 rRNA y 22 tRNA.
- El DNA mitocondrial de levadura es 5 veces más largo que el mtDNA de una célula animal debido a la presencia de intrones largos.

REVISIÓN DE CONCEPTOS

¿Qué explica la variación en el tamaño de los genomas mitocondriales entre diferentes eucariotes?

4.11 El genoma de cloroplasto codifica para muchas proteínas y para RNA

¿Qué genes portan los cloroplastos? Los DNA de cloroplastos varían en longitud desde ~120 a 217 kb (el más largo en el geranio). Los genomas de cloroplastos secuenciados (> 100 en total) tienen entre 87 y 183 genes. La **FIGURA 4.19** resume las funciones codificadas por el genoma de cloroplasto en las plantas terrestres. Existe más variación en el genoma de los cloroplastos de las algas.

FIGURA 4.19 El genoma de cloroplasto en las plantas terrestres codifica 4 rRNA, 30 tRNA y ~60 proteínas.

| Genes | Tipos |
|---------------------------|---------|
| RNA que codifica | |
| 16S rRNA | 1 |
| 23S rRNA | 1 |
| 4.5S rRNA | 1 |
| 5S rRNA | 1 |
| tRNA | 30–32 |
| Expresión del gen | |
| proteínas r | 20–21 |
| RNA polimerasa | 3 |
| Otras | 2 |
| Funciones del cloroplasto | |
| Rubisco y tilacoides | 31–32 |
| NADH deshidrogenasa | 11 |
| Total | 105–113 |

El genoma del cloroplasto es generalmente similar al de la mitocondria, excepto que involucra más genes. El genoma de cloroplasto codifica para todas las especies de rRNA y tRNA necesarias para la síntesis de proteína. El ribosoma incluye dos rRNA pequeños, además de las especies principales. El grupo de tRNA puede incluir todos los genes necesarios. El genoma de cloroplasto codifica para ~50 proteínas, y también para la RNA polimerasa y las proteínas ribosómicas. Nuevamente, la regla es que los genes de los orgánulos se transcriben y traducen dentro del orgánulo. Alre-

dedor de la mitad de los genes de cloroplasto codifican proteínas involucradas en la síntesis proteica.

Los intrones en los cloroplastos pertenecen a dos clases generales. Aquellos en los genes de tRNA suelen estar (aunque no inevitablemente) localizados en el bucle anticodón, al igual que los intrones que se encuentran en los genes de tRNA nuclear de levadura (véase la Sección 28.13, *El corte y empalme del tRNA de la levadura consiste en corte y la reunión*). Aquellos en los genes que codifican proteína se asemejan a los intrones de los genes mitocondriales (véase el Capítulo 29, *RNA catalítico*). Esto coloca al acontecimiento endosimbiótico en un tiempo en la evolución antes de la separación de los procariontes con genes sin interrupciones.

El cloroplasto es el sitio de la fotosíntesis. Muchos de sus genes codifican las proteínas de los complejos fotosintéticos localizados en las membranas tilacoides. La constitución de estos complejos muestra una proporción diferente al de los complejos mitocondriales. Aunque algunos complejos son similares a los de las mitocondrias, en que tienen algunas subunidades codificadas por el genoma del orgánulo y algunas por el genoma nuclear, otros complejos de cloroplastos están codificados enteramente por un genoma. Por ejemplo, el gen para la subunidad grande de la ribulosa bifosfato carboxilasa (RuBisCO, que cataliza la reacción de fijación del carbono del ciclo de Calvin), *rbcL*, está contenido en el genoma del cloroplasto; la variación en este gen es utilizada frecuentemente como base para la reconstrucción de las filogenias de plantas. Sin embargo, el gen para la subunidad pequeña de la rubisco, *rbcS*, en general suele portarlo el genoma nuclear. Por el otro lado, los genes para los complejos proteicos del fotosistema se encuentran en el genoma del cloroplasto, mientras que aquellos para las proteínas del LHC (complejo de captación de luz) están codificados en el núcleo.

CONCEPTOS CLAVE

- Los genomas de cloroplasto varían en tamaño, pero son los suficientemente grandes para codificar de 50 a 100 proteínas, así como los rRNA y tRNA.

REVISIÓN DE CONCEPTOS

¿Qué ventajas existen en los genomas de cloroplasto que codifican la RNA polimerasa, los tRNA, los rRNA y las proteínas ribosómicas?

4.12 Los cloroplastos y las mitocondrias evolucionaron por endosimbiosis

¿Cómo fue que un orgánulo haya evolucionado de modo tal como para contener información genética para algunas de sus funciones, mientras que la información para otras funciones está codificada en el núcleo? La FIGURA 4.20 muestra la hipótesis endosimbiótica para la evolución mitocondrial, en la cual las células primitivas capturaron bacterias que proporcionaron la función de la respiración celular, las que con el transcurso del tiempo

evolucionaron hasta convertirse en mitocondrias. En este punto, el protorgánulo debe haber contenido todos los genes necesarios para especificar sus funciones. Para explicar el origen de los cloroplastos, se propuso un mecanismo similar.

Las homologías de secuencias sugieren que las mitocondrias y los cloroplastos evolucionaron de forma separada, a partir de linajes que son comunes con diferentes eubacterias, las mitocondrias que comparten un origen con las bacterias púrpuras alfa y los cloroplastos que comparten un origen con las cianobacterias. El pariente más cercano conocido de las mitocondrias entre las bacterias es *Rickettsia* (el agente causante del tifus), que es un parásito intracelular obligado que probablemente descendió de bacterias de vida libre. Esto refuerza la idea de que las mitocondrias se originaron en un acontecimiento endosimbiótico que involucra a un ancestro que también es común a *Rickettsia*.

El origen endosimbiótico del cloroplasto está enfatizado por las relaciones entre sus genes y sus contrapartes en las bacterias. La organización de los genes del rRNA en particular se relaciona estrechamente con la organización de una cianobacteria, que se atribuye más precisamente al último ancestro común entre cloroplastos y bacterias. No es sorprendente que las cianobacterias sean fotosintéticas.

A medida que las bacterias se integraron dentro de la célula receptora y evolucionaron para convertirse en la mitocondria (o cloroplasto), debieron haberse producido dos cambios. Los orgánulos tienen menos genes que una bacteria independiente y han perdido muchas de las funciones génicas que son necesarias para la vida independiente (como las vías metabólicas). La mayoría de los genes que codifican las funciones de los orgánulos se localiza, en la actualidad, en el núcleo; por lo tanto, estos genes deben haber sido transferidos allí desde el orgánulo.

La transferencia de DNA entre un orgánulo y el núcleo ha ocurrido durante la historia evolutiva y aún continúa. La tasa de transferencia puede medirse directamente al introducir dentro de un orgánulo un gen que puede funcionar solamente en el núcleo; por ejemplo, debido a que éste contiene un intrón nuclear, o debido a que la proteína debe funcionar en el citosol. En términos de provisión de material para la evolución, las tasas de transferencia desde el orgánulo al núcleo son apenas equivalentes a la tasa de mutación de un gen único. El DNA introducido dentro de la mitocondria es transferido al núcleo a una tasa de 2×10^{-5} por generación. Los experimentos para medir la transferencia en la dirección inversa, del núcleo a la mitocondria, sugieren que la tasa es mucho menor, $< 10^{-10}$. Cuando se introduce un gen de resistencia a un antibiótico específico nuclear dentro de cloroplastos, su transferencia al núcleo y el éxito de la expresión pueden seguirse mediante la detección de plántulas resistentes al antibiótico. Esto demuestra que la transferencia ocurre a una tasa de 1 en 16.000 plántulas, o 6×10^{-5} por generación.

La transferencia de un gen de un orgánulo al núcleo requiere movimiento físico del DNA, por supuesto, pero la expresión exitosa también requiere cambios en la secuencia codificante. Las proteínas de orgánulos que están codificadas por genes nucleares tienen secuencias especiales que les permiten ser importadas dentro del orgánulo después de que han sido sintetizadas en el citoplasma (véase la Sección 10.6, *La inserción postraducciona en la membrana depende de secuencias señal*). Estas secuencias no son requeridas por las proteínas que se sintetizan dentro del orgánulo. Tal vez, los procesos de transferencia génica efectiva

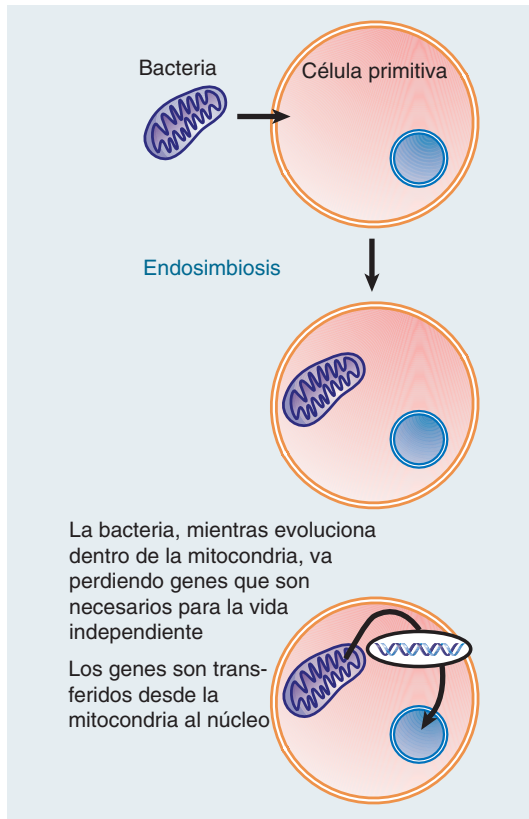


FIGURA 4.20 La mitocondria se originó a partir de un acontecimiento endosimbiótico cuando una célula eucarionte capturó una bacteria.

ocurrieron en un período cuando los compartimentos estuvieron menos estrictamente definidos, de manera tal que fue más fácil tanto para el DNA ser relocalizado como para las proteínas ser incorporadas dentro del orgánulo independientemente del sitio de síntesis.

Los mapas filogenéticos muestran que las transferencias génicas ocurrieron independientemente en muchos linajes diferentes. Parece ser que las transferencias de genes mitocondriales al núcleo se produjeron sólo al comienzo de la evolución de la célula animal, pero es posible que el proceso aún continúe en las células vegetales. El número de transferencias puede ser grande: existen 800 genes nucleares en *Arabidopsis* cuyas secuencias están relacionadas con los genes en los cloroplastos de otras plantas. Estos genes son candidatos para la evolución a partir de los genes que se originaron en el cloroplasto.

CONCEPTOS CLAVE

- Tanto las mitocondrias como los cloroplastos descienden de ancestros bacterianos.
- La mayoría de los genes de los genomas mitocondriales y de cloroplastos han sido transferidos al núcleo durante la evolución del orgánulo.

REVISIÓN DE CONCEPTOS

1. ¿Qué evidencia existe de que los cloroplastos descienden de ancestros procariontes?
2. Diseña un experimento que mida la tasa de transferencia de genes desde las mitocondrias al núcleo en células de levadura.

4.13 Resumen

Las secuencias de DNA que componen un genoma eucarionte pueden clasificarse en tres grupos:

- secuencias no repetitivas que son únicas;
- secuencias moderadamente repetitivas que se encuentran dispersas y repetidas una cantidad pequeña de veces con algunas copias que no son idénticas; y
- secuencias altamente repetitivas que son cortas y que suelen estar repetidas y organizadas en tándem.

Las proporciones de estos tipos de secuencias son características de cada genoma, aunque los genomas más grandes tienden a tener una menor proporción de DNA no repetitivo. Casi el 50% del genoma humano consiste en secuencias repetitivas –la vasta mayoría corresponden a las secuencias de transposones. La mayoría de los genes estructurales se localiza en el DNA no repetitivo. La cantidad de DNA no repetitivo refleja mejor la complejidad del organismo que el tamaño total del genoma; la mayor cantidad de DNA no repetitivo en los genomas es $\sim 2 \times 10^9$ pb.

La herencia no mendeliana se explica mediante la presencia de DNA en los orgánulos en el citoplasma. Las mitocondrias y los cloroplastos son sistemas rodeados por membranas en los cuales algunas proteínas se sintetizan dentro del orgánulo mientras que otras se importan. El genoma del orgánulo es, usualmente, un DNA circular que codifica todos los RNA y algunas proteínas requeridas por el orgánulo.

Los genomas mitocondriales varían mucho en tamaño desde el genoma pequeño de 16 kb en mamíferos hasta el genoma de 570 kb en las plantas superiores. Los genomas más grandes pueden codificar funciones adicionales. Los genomas de cloroplastos varían desde ~ 120 a 217 kb. Aquellos que se han secuenciado tienen una organización similar y codifican funciones equivalentes. Tanto en las mitocondrias como en los cloroplastos, muchas de las proteínas más importantes contienen algunas subunidades sintetizadas en el orgánulo y algunas subunidades importadas desde el citosol. Se han producido transferencias de DNA entre los genomas de cloroplastos o de mitocondrias, y el nuclear.

PREGUNTAS

1. ¿Aproximadamente con qué frecuencia se producen polimorfismos de un nucleótido único (SNP) en el genoma humano?
 - A. uno cada 560 bases
 - B. uno cada 950 bases
 - C. uno cada 1.330 bases
 - D. uno cada 2.040 bases
2. Si una cruce genética entre dos moscas –una con pigmento oscuro del ojo y un fragmento de DNA único que hibridiza con una sonda y otra que carece de pigmento y tiene dos fragmentos que hibridizan con la sonda– produce una progenie con un 40% de un tipo parental y un 40% del otro tipo parental, y un 10% de cada fenotipo recombinante, ¿cuál es la distancia en el mapa genético entre el marcador de restricción y el marcador del color de ojo?
 - A. 10%
 - B. 20%
 - C. 40%
 - D. 80%
3. ¿Cuál de los siguientes grupos de organismos exhibe la menor complejidad del genoma?
 - A. insectos
 - B. moluscos
 - C. hongos
 - D. algas
4. ¿Cuál especie exhibe el mayor porcentaje de DNA repetitivo?
 1. *C. elegans*
 2. *X. laevis*
 3. *D. melanogaster*
 4. *H. sapiens*
5. Los genomas procariontes suelen contener:
 - A. casi el 100% de DNA no repetitivo
 - B. mayormente DNA no repetitivo con alrededor de un 20% de DNA moderadamente repetitivo.
 - C. mayormente DNA moderada o altamente repetitivo con alrededor de un 20% de DNA no repetitivo.
 - D. alrededor de cantidades iguales de DNA moderadamente repetitivo y no repetitivo.
6. Los genomas de plantas y anfibios contienen:
 - A. casi el 100% de DNA no repetitivo.
 - B. mayormente DNA no repetitivo con alrededor de un 20% de DNA moderadamente repetitivo.
 - C. mayormente DNA moderada o altamente repetitivo con alrededor de un 20% de DNA no repetitivo.
 - D. alrededor de cantidades iguales de DNA moderadamente repetitivo y no repetitivo.
7. La distrofia muscular de *Duchenne* (DMD) es causada por un(a) _____ en el gen de la distrofina:
 - A. delección grande
 - B. inserción grande
 - C. cambio de base único
 - D. inserción pequeña.
8. En la mayoría de los animales, el DNA mitocondrial es heredado
 - A. de igual manera de ambos padres.
 - B. predominantemente del padre.
 - C. sólo de la madre.
 - D. sólo del padre.

9. Durante el curso de la evolución, los genes de los genomas mitocondriales y de cloroplastos han:
 - A. sido transferidos a los cromosomas nucleares.
 - B. sido transferidos desde un orgánulo al otro.
 - C. sido eliminados de la célula.
 - D. evolucionado a otras funciones.
10. El pariente más cercano conocido de la mitocondria entre las bacterias es:
 - A. cianobacteria.
 - B. *Rickettsia*.
 - C. *Salmonella*.
 - D. *Bacillus*.

PALABRAS CLAVE

| | | | |
|------------------------|------------------------|-----------------------------|------------------------------|
| ctDNA (cpDNA) | mapa genético | proteoma | DNA redundante |
| valor C | genoma | DNA repetitivo | polimorfismo de un |
| paradoja del valor C | haplotipo | polimorfismo de longitud en | nucleótido único (SNP) |
| paseo cromosómico | interactoma | los fragmentos de | sintenia |
| bucle D | mapa de ligamiento | restricción (RFLP) | transcriptoma |
| huella genética | herencia materna | mapa de restricción | transposón |
| atrapamiento de exones | mtDNA | transcripción inversa- | transferencia zoo (zoo blot) |
| etiqueta de secuencia | herencia no mendeliana | reacción en cadena de la | |
| expresada (EST) | DNA no repetitivo | polimerasa (RT-PCR) | |
| genes extranucleares | polimorfismo | | |

LECTURAS RECOMENDADAS

- Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., and Donnelly, P. (2005). A haplotype map of the human genome. *Nature* 437, 1299–1320.
- Un informe sobre *HapMap*, la base de datos pública de los SNP (polimorfismo de un nucleótido único), que representa un intento a gran escala de documentar la variación genética humana.
- Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101.
- Una revisión del enigma (paradoja) del valor C y de la hipótesis que la explica, y un análisis de la correlación positiva entre el valor C y el tamaño de la célula.
- Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079.
- Un informe acerca de 1,5 millones de SNP de 71 individuos humanos que descendieron a partir de tres poblaciones diferentes como una muestra de la variación genética humana.
- Lang, B. F., Gray, M. W., and Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33, 351–397.
- La evidencia sugiere que las mitocondrias evolucionaron a partir de una α -proteobacteria endosimbiótica alrededor de la misma época que evolucionó el núcleo.
- Sharp, A. J., Cheng, Z., and Eichler, E. E. (2006). Structural variation of the human genome. *Annu. Rev. Genom. Hum. G.* 7, 407–442.
- Una revisión de los reordenamientos estructurales del genoma humano y sus efectos en la variación fenotípica.
- Sugiura, M., Hirose, T., and Sugita, M. (1998). Evolution and mechanism of translation in chloroplasts. *Annu. Rev. Genom. Hum. G.* 32, 437–459.