9

Tecnologias da Informação com Base no DNA

A complexidade das moléculas e dos sistemas apresentados neste livro pode, algumas vezes, ocultar uma realidade bioquímica – o que se aprendeu é apenas o começo. Novos lipídeos, proteínas, carboidratos e ácidos nucleicos são descobertos a cada dia e, com frequência, não se tem nenhuma ideia sobre suas funções. Quantos ainda serão descobertos e quais serão suas funções? Até mesmo moléculas bem caracterizadas continuam a desafiar os pesquisadores com inúmeras questões mecânicas e funcionais não resolvidas. Os milhares de sequências genômicas completas já conhecidas fornecem uma estimativa da imensidão da tarefa à frente. Simplesmente não são conhecidas as funções da maior parte do DNA – em geral, incluindo a metade ou mais dos genes – em um genoma típico. Essas mesmas sequências genômicas também fornecem uma oportunidade sem precedentes. Não há maior fonte de informações sobre uma célula ou organismo do que as inseridas em seu próprio DNA. A tarefa de extrair essas informações originou sofisticadas tecnologias com base no DNA que agora estão no centro de quase toda a história da bioquímica. Essas tecnologias serão abordadas em cada um dos tópicos explorados nos próximos capítulos.

Como objetos de estudo, as moléculas de DNA apresentam um problema especial – seu tamanho. Sem dúvida, os cromossomos são as maiores biomoléculas na célula. Como um pesquisador descobre a informação que procura quando só uma pequena parte de um cromossomo inclui milhões ou até bilhões de pares de bases contíguos? Soluções para esses problemas começaram a surgir na década de 1970.

Décadas de avanços por milhares de cientistas que trabalham em genética, bioquímica, biologia celular e físico-química se uniram nos laboratórios de Paul Berg, Herbert Boyer e Stanley Cohen para produzir técnicas de detecção, isolamento, preparação e estudo de pequenos segmentos de DNA derivados de cromossomos muito maiores. As técnicas para a clonagem de DNA pavimentaram o caminho para o moderno campo da **genômica** e, de forma mais ampla, para muitas das tecnologias que contribuem para a **biologia de sistemas**, o estudo da bioquímica na escala de todas as células e organismos.

Cada aluno ou instrutor, quando considera os temas apresentados neste capítulo, depara-se com um conflito. Primeiro, os métodos aqui descritos foram possíveis graças aos avanços na compreensão do metabolismo do DNA e do RNA. Assim, é preciso entender alguns conceitos fundamentais da replicação do DNA, transcrição do RNA, síntese de proteínas e regulação gênica para entender como esses métodos funcionam. Ao mesmo tempo, no entanto, a bioquímica moderna se baseia nesses mesmos métodos, de modo que uma abordagem atual sobre qualquer aspecto da disciplina torna-se muito difícil sem uma introdução adequada a eles. Apresentar esses métodos no início do livro é reconhecer que eles estão intrinsecamente interligados tanto aos avanços que lhes deram origem quanto às mais novas descobertas que eles hoje tornam possíveis. Esses conceitos básicos e necessários fazem da presente discussão não apenas uma introdução à tecnologia, mas também uma prévia de muitos dos fundamentos da bioquímica do DNA e do RNA encontrados em capítulos posteriores.

Inicialmente são destacados os princípios de clonagem do DNA e, em seguida, ilustrados a variedade de aplicações

Paul Berg

Herbert Boyer

Stanley N. Cohen

e o potencial de muitas das tecnologias mais recentes que apoiam e aceleram o avanço da bioquímica.

9.1 Estudo dos genes e seus produtos

Uma pesquisadora isolou uma nova enzima que é a chave para uma doença humana. Ela espera isolar grandes quantidades da proteína a fim de cristalizá-la para seu estudo e análise estrutural. Deseja alterar os resíduos de aminoácidos em seu sítio ativo para compreender a reação que a enzima catalisa. Ela prepara um sofisticado programa de pesquisa para elucidar como essa enzima interage e é regulada por outras proteínas na célula. Tudo isso, e muito mais, se torna possível se ela conseguir obter o gene que codifica sua enzima. Infelizmente, esse gene tem apenas alguns milhares de pares de bases no interior de um cromossomo humano, com um tamanho medido em centenas de milhares de pares de bases. Como ela vai isolar o pequeno segmento de que necessita e, então, estudá-lo? A resposta encontra-se na clonagem do DNA e em métodos desenvolvidos para manipular genes clonados.

Genes podem ser isolados por clonagem do DNA

Um *clone* é uma cópia idêntica. Esse termo foi originalmente aplicado a células de um único tipo, isoladas e cultivadas para criar uma população de células idênticas. Quando aplicado ao DNA, um clone representa muitas cópias idênticas de um segmento particular do gene. Em resumo, a pesquisadora deve cortar o gene do cromossomo maior, anexá-lo a uma porção muito menor de DNA transportador e permitir que microrganismos façam muitas cópias dessa porção para ela. Esse é o processo de **clonagem de DNA**. O resultado é a amplificação seletiva de um determinado gene ou segmento de DNA, facilitando seu isolamento e estudo. Classicamente, a clonagem de DNA a partir de qualquer organismo envolve cinco procedimentos gerais:

1. *Cortar o DNA-alvo em locais precisos*. Endonucleases específicas de sequência (endonucleases de restrição) fornecem as tesouras moleculares necessárias.
2. *Selecionar uma pequena molécula transportadora de DNA capaz de autorreplicação*. Esses DNAs são chamados **vetores de clonagem** (vetores são agentes de entrega). São geralmente plasmídeos ou DNAs virais.
3. *Juntar dois fragmentos de DNA de modo covalente*. A enzima DNA-ligase liga o vetor de clonagem e o DNA a ser clonado. Moléculas de DNA compostas por segmentos ligados de modo covalente a partir de duas ou mais fontes são chamadas de **DNAs recombinantes**.
4. *Transportar o DNA recombinante do tubo de ensaio para uma célula hospedeira* que irá proporcionar a maquinaria enzimática para a replicação do DNA.
5. *Selecionar ou identificar as células hospedeiras que contêm DNA recombinante.*

Os métodos utilizados para realizar essas e outras tarefas relacionadas são coletivamente chamados de **tecnologia do DNA recombinante** ou, de modo mais informal, de **engenharia genética**.

Grande parte dessa discussão inicial incidirá sobre a clonagem de DNA na bactéria *Escherichia coli*, o primeiro organismo utilizado para o trabalho de DNA recombinante e ainda a célula hospedeira mais comum. A *E. coli* apresenta muitas vantagens: o metabolismo do seu DNA (como muitos outros de seus processos bioquímicos) é bem compreendido; muitos vetores de clonagem que ocorrem naturalmente associados à *E. coli*, como plasmídeos e bacteriófagos (vírus bacterianos, também chamado de fagos), são bem caracterizados; e existem técnicas para transportar o DNA rapidamente de uma célula bacteriana a outra. Os princípios aqui discutidos são amplamente aplicáveis à clonagem de DNA em outros organismos, tópico discutido em mais detalhes adiante na seção.

Endonucleases de restrição e DNA-ligases produzem DNA recombinante

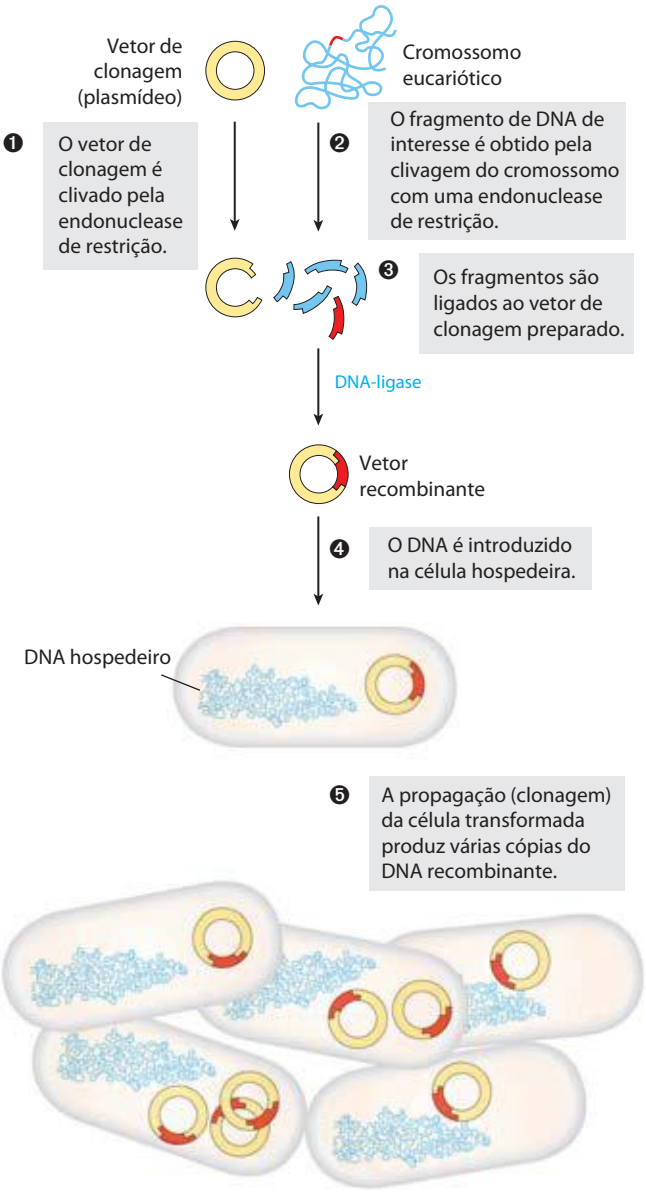
Particularmente importante para a tecnologia do DNA recombinante é um conjunto de enzimas (Tabela 9-1) que se tornou disponível ao longo de décadas de pesquisa sobre o metabolismo de ácidos nucleicos. Duas classes de enzimas estão no centro da abordagem clássica para a geração e propagação da molécula de DNA recombinante (**Figura 9-1**). Em primeiro lugar, as **endonucleases de restrição** (também chamadas de enzimas de restrição) reconhecem e clivam o DNA em sequências específicas (sequências de reconhecimento ou sítios de restrição) para gerar um conjunto de fragmentos menores. Em segundo lugar, o fragmento do DNA a ser clonado é unido a um vetor de clonagem adequado, por meio de **DNA-ligases**, para ligar as moléculas de DNA. O vetor recombinante é então introduzido em uma célula hospedeira que amplifica o fragmento no curso de muitas gerações de divisão celular.

Endonucleases de restrição são encontradas em um grande número de espécies bacterianas. No início da década de 1960, Werner Arber descobriu que a função biológica dessas enzimas consiste em reconhecer e clivar o DNA exógeno (o DNA de um vírus infeccioso, por exemplo); afirma-se que tal DNA está *restrito*. No DNA da célula hospedeira, a sequência que poderia ser reconhecida por sua própria endonuclease de restrição é protegida da digestão pela metilação do DNA, catalisada por uma metilase específica. A endonuclease de restrição e a metilase correspondente são chamadas algumas vezes de **sistema de restrição-modificação**.

Há três tipos de endonucleases de restrição, designadas I, II e III. Os tipos I e III são complexos geralmente grandes com múltiplas subunidades, contendo atividades tanto de endonuclease quanto de metilase. As endonucleases de restrição tipo I clivam o DNA em sítios aleatórios que podem estar a mais de 1.000 pares de base (pb) da sequência de reconhecimento. As endonucleases de restrição tipo III clivam o DNA a cerca de 25 pb da sequência de reconhecimento. Ambos os tipos se movem ao longo

TABELA 9-1 Algumas enzimas usadas na tecnologia do DNA recombinante

Enzima(s)	Função
Endonucleases de restrição tipo II	Clivam DNA em sequências de bases específicas
DNA-ligase	Liga duas moléculas ou fragmentos de DNA
DNA-polimerase I (*E. coli*)	Preenche lacunas em DNA de fita dupla com adição de nucleotídeos às extremidades 3′
Transcriptase reversa	Faz uma cópia de DNA a partir de uma molécula de RNA
Polinucleotídeo-cinase	Adiciona um fosfato à extremidade 5′-OH de um polinucleotídeo para marcá-la ou permitir ligação
Terminal-transferase	Adiciona caudas de homopolímeros às extremidades 3′-OH de um duplex linear
Exonuclease III	Remove resíduos nucleotídicos das extremidades 3′ de uma fita de DNA
Exonuclease do bacteriófago λ	Remove resíduos nucleotídicos das extremidades 5′ de um duplex de DNA para expor as extremidades 3′ de fita simples
Fosfatase alcalina	Remove fosfatos terminais das extremidades 5′ou 3′ (ou ambas)



do DNA em uma reação que precisa da energia do ATP. As **endonucleases de restrição tipo II**, isoladas pela primeira vez por Hamilton Smith em 1970, são mais simples, não requerem ATP e catalisam a clivagem hidrolítica de ligações fosfodiésteres específicas no DNA dentro da própria sequência de reconhecimento. A extraordinária utilidade desse grupo de endonucleases de restrição foi demonstrada por Daniel Nathans, que primeiro as utilizou para desenvolver novos métodos de mapeamento e análise de genes e genomas.

Milhares de endonucleases de restrição tipo II foram descobertos em diferentes espécies bacterianas, e mais de 100 sequências de DNA diferentes são reconhecidas por uma ou mais dessas enzimas. As sequências de reconhecimento têm em geral 4 a 6 pb de comprimento e são palindrômicas (ver Figura 8-18). A Tabela 9-2 lista as sequências reconhecidas por algumas endonucleases tipo II.

Algumas endonucleases de restrição fazem cortes coordenados nas duas fitas do DNA, deixando dois a quatro nucleotídeos de uma fita sem par em cada extremidade resultante. Essas fitas sem pares são chamadas de **extremidades adesivas** (**Figura 9-2a**), pois formam pares de bases entre si ou com extremidades adesivas complementares de outros fragmentos de DNA. Outras endonucleases de restrição clivam ambas as fitas do DNA nas ligações fosfodiésteres opostas, sem deixar nenhuma base sem par nas extremidades, frequentemente chamadas de **extremidades cegas** (Figura 9-2b).

O tamanho médio dos fragmentos de DNA produzidos pela clivagem do DNA genômico com uma endonuclease de

FIGURA 9-1 Ilustração esquemática da clonagem do DNA. Um vetor de clonagem e cromossomos eucarióticos são clivados separadamente com a mesma endonuclease de restrição (um único cromossomo é mostrado para fins de simplificação). Os fragmentos a serem clonados são então ligados ao vetor de clonagem. O DNA recombinante resultante (apenas um vetor recombinante é mostrado aqui) é introduzido em uma célula hospedeira, onde pode ser propagado (clonado). Observe que este desenho não está em escala: o tamanho do cromossomo de *E. coli* em relação àquele de um típico vetor de clonagem (como um plasmídeo) é muito maior do que o mostrado aqui.

TABELA 9-2 Sequências de reconhecimento para algumas endonucleases de restrição tipo II

*Bam*HI	(5′) G G A T C C (3′) C C T A G G	*Hin*dIII	(5′) A A G C T T (3′) T T C G A A
*Cla*I	(5′) A T C G A T (3′) T A G C T A	*Not*I	(5′) G C G G C C G C (3′) C G C C G G C G
*Eco*RI	(5′) G A A T T C (3′) C T T A A G	*Pst*I	(5′) C T G C A G (3′) G A C G T C
*Eco*RV	(5′) G A T A T C (3′) C T A T A G	*Pvu*II	(5′) C A G C T G (3′) G T C G A C
*Hae*III	(5′) G G C C (3′) C C G G	*Tth*111I	(5′) G A C N N N G T C (3′) C T G N N N C A G

Nota: As setas indicam as ligações de fosfodiéster clivadas por cada endonuclease de restrição. Os asteriscos indicam bases metiladas pela metilase correspondente (quando conhecida). N denota qualquer base. Observe que o nome de cada enzima consiste em uma abreviação de três letras (em itálico) da espécie bacteriana da qual ela deriva, por vezes seguida por designação da linhagem e de números romanos para distinguir diferentes endonucleases de restrição isoladas das mesmas espécies bacterianas. Então *Bam*H I é a primeira (I) endonuclease de restrição caracterizada isolada da linhagem H do *B*acyllus *amy*loliquefaciens.

restrição depende da frequência na qual um sítio específico de restrição ocorre na molécula do DNA, que, por sua vez, depende muito do tamanho da sequência de reconhecimento.

Em uma molécula de DNA com sequência aleatória na qual todos os quatro nucleotídeos são igualmente abundantes, uma sequência de 6 pb reconhecida por uma endonuclease de restrição como *Bam*HI pode ocorrer, em média, uma vez a cada 4^6 (4.096) pb. Enzimas que reconhecem uma sequência de 4 pb produzem fragmentos de DNA menores a partir de uma molécula de DNA com sequência aleatória; uma sequência de reconhecimento desse tamanho ocorreria mais ou menos a cada 4^4 (256) pb. Em moléculas naturais de DNA, sequências de reconhecimento específicas tendem a ocorrer com menos frequência do que essas, porque as sequências nucleotídicas no DNA não são aleatórias e os quatro nucleotídeos não são igualmente abundantes. Em experimentos laboratoriais, o tamanho médio de fragmentos produzidos pela clivagem por endonucleases de restrição de um DNA grande pode ser aumentado simplesmente terminando a reação antes de se completar o processo; o resultado é chamado de digestão parcial. O tamanho médio do fragmento também pode ser aumentado utilizando-se uma classe especial de endonucleases, chamadas de endonucleases *homing* (ver Figura 26-37). Elas reconhecem e clivam sequências de DNA muito maiores (14 a 20 pb).

Uma vez que a molécula de DNA tenha sido clivada em fragmentos, um determinado fragmento de tamanho conhecido pode ser parcialmente purificado por meio de eletroforese em gel de agarose ou poliacrilamida (p. 302) ou por HPLC (p. 92). Para um genoma típico de mamífero, entretanto, a clivagem por endonucleases de restrição geralmente produz um número muito grande de fragmentos de DNA diferentes para permitir o isolamento adequado de um segmento específico. Uma etapa intermediária comum na clonagem de um gene específico ou segmento de DNA é a construção de uma biblioteca de DNA (descrita na Seção 9.2).

Após o isolamento do fragmento de DNA-alvo, a DNA-ligase pode ser utilizada para uni-lo a um vetor de clonagem digerido de modo semelhante –isto é, um vetor digerido pela *mesma* endonuclease de restrição; um fragmento gerado por *Eco*RI, por exemplo, em geral não irá se unir a um fragmento gerado por *Bam*HI. Como descrito mais detalhadamente no Capítulo 25 (ver Figura 25-16), a DNA-ligase catalisa a formação de novas ligações fosfodiésteres em uma reação que utiliza ATP ou um cofator semelhante. O pareamento de bases das extremidades adesivas complementares facilita muito a reação da ligação (Figura 9-2a). As extremidades cegas também podem ser ligadas, embora com menos eficiência. Os pesquisadores podem criar novas sequências de DNA inserindo fragmentos sintéticos de DNA (chamados de ***linkers***) entre as extremidades que estão sendo ligadas. Fragmentos de DNA inseridos com múltiplas sequências de reconhecimento para endonucleases de restrição (frequentemente úteis como pontos para inserção posterior de DNA adicional por clivagem e ligação) são chamados de ***polilinkers*** (Figura 9-2c).

A eficácia das extremidades adesivas em unir seletivamente dois fragmentos de DNA foi revelada nos primeiros experimentos de DNA recombinante. Antes que as endonucleases estivessem amplamente disponíveis, alguns pesquisadores descobriram que podiam gerar extremidades adesivas pela ação combinada da exonuclease do bacteriófago λ e da enzima terminal-transferase (Tabela 9-1). Eram adi-

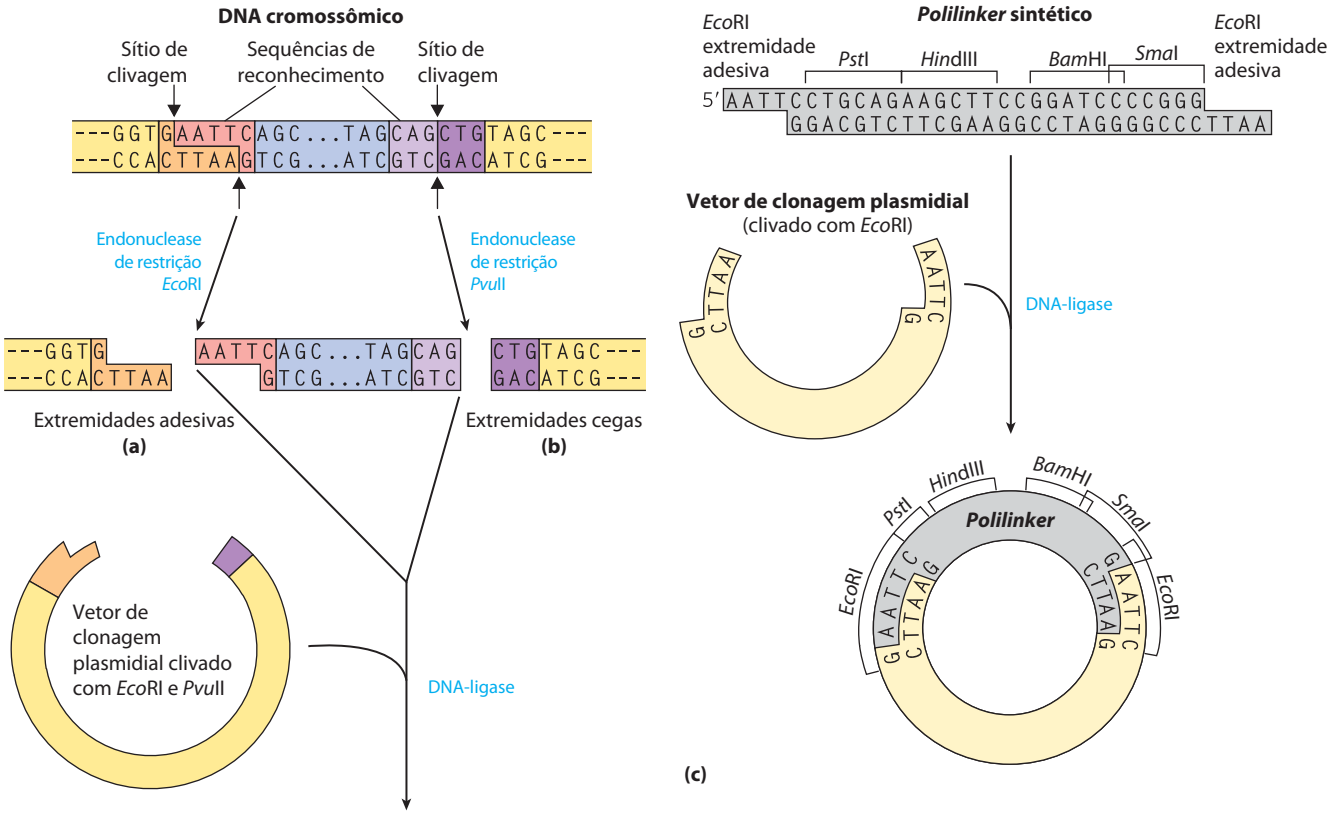


FIGURA 9-2 Clivagem de moléculas de DNA por endonucleases de restrição. As endonucleases de restrição reconhecem e clivam apenas sequências específicas, deixando tanto as extremidades adesivas **(a)** (com fitas simples salientes) quanto as extremidades cegas **(b)**. Os fragmentos podem se ligar a outros DNA, como o vetor de clonagem clivado (um plasmídeo) mostrado aqui. Esta reação é facilitada pela renaturação das extremidades adesivas complementares. A ligação é menos eficiente para os fragmentos de DNA com extremidades cegas do que para aqueles com extremidades adesivas complementares, e os fragmentos de DNA com diferentes extremidades adesivas (não complementares) geralmente não se ligam. **(c)** Um fragmento de DNA sintético com sequências de reconhecimento para várias endonucleases de restrição pode ser inserido em um plasmídeo que foi clivado por uma endonuclease de restrição. A inserção é chamada de *linker*; uma inserção com múltiplos sítios de restrição é chamada de *polilinker*. **Endonucleases de restrição**

cionadas caudas homopoliméricas aos fragmentos a serem unidos. Peter Lobban e Dale Kaiser utilizaram esse método em 1971 nos primeiros experimentos para unir fragmentos de DNA de ocorrência natural. Métodos semelhantes foram utilizados logo após no laboratório de Paul Berg para unir segmentos de DNA do vírus de símios 40 (SV40) ao DNA derivado de bacteriófago λ, criando assim a primeira molécula de DNA recombinante com segmentos de DNA de espécies diferentes.

Os vetores de clonagem permitem a amplificação dos segmentos de DNA inseridos

Os princípios que regem a liberação de DNA recombinante em uma forma clonável a uma célula hospedeira e sua posterior amplificação nessa célula são bem ilustrados considerando-se dois vetores de clonagem comuns – os plasmídeos e os cromossomos artificiais de bactérias – utilizados em experimentos com *E. coli* e um vetor usado para clonar segmentos de DNA grandes em levedura.

Plasmídeos Um plasmídeo é uma molécula de DNA circular que se replica separadamente do cromossomo hospedeiro. A grande variedade de plasmídeos bacterianos que ocorre naturalmente varia de tamanho de 5.000 a 400.000 pb. Muitos dos plasmídeos encontrados em populações bacterianas são um pouco mais do que parasitas moleculares, semelhantes aos vírus, mas com capacidade mais limitada de se transferir de uma célula para outra. Para sobreviver na célula hospedeira, os plasmídeos incorporam várias sequências especializadas que os tornam capazes de utilizar a energia das células para sua própria replicação e expressão gênica.

Os plasmídeos de ocorrência natural têm um papel simbiótico na célula: são capazes de fornecer genes que conferem resistência a antibióticos ou que realizam novas funções para a célula. Por exemplo, o plasmídeo Ti de *Agro-*

bacterium tumefaciens torna a bactéria hospedeira capaz de colonizar as células de plantas e se utilizar de sua energia. As mesmas propriedades que tornam os plasmídeos capazes de sobreviverem em um hospedeiro bacteriano ou eucariótico são úteis para os biólogos moleculares desenvolverem um vetor para a clonagem de um segmento específico de DNA. O clássico plasmídeo de *E. coli* pBR322, desenvolvido em 1977, é um bom exemplo de um plasmídeo com características úteis em quase todos os vetores de clonagem (**Figura 9-3**):

1. O plasmídeo pBR322 tem uma **origem de replicação**, ou **ori**, uma sequência onde a replicação é iniciada por enzimas celulares (ver Capítulo 25). Essa sequência é necessária para propagar o plasmídeo. Um sistema regulatório associado limita a replicação para manter o pBR322 em um nível de 10 a 20 cópias por célula.
2. O plasmídeo contém genes que conferem resistência aos antibióticos tetraciclina (Tet^R) e ampicilina (Amp^R), permitindo a seleção de células que contêm o plasmídeo intacto ou uma versão recombinante do plasmídeo (discutida a seguir).
3. Várias sequências de reconhecimento únicas no pBR322 são alvo para endonucleases de restrição (*Pst*I, *Eco*RI, *Bam*HI, *Sal*I e *Pvu*II), fornecendo sítios onde o plasmídeo pode ser cortado e inserido em um DNA exógeno.
4. O pequeno tamanho do plasmídeo (4.361 pb) facilita sua entrada nas células e a manipulação bioquímica do DNA. Esse pequeno tamanho foi gerado simplesmente cortando-se muitos segmentos de DNA a partir de um plasmídeo parental maior – sequências que o biólogo molecular não necessita.

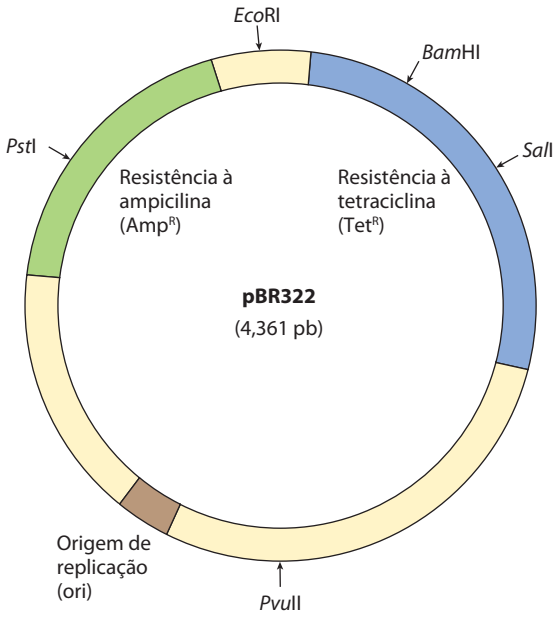


FIGURA 9-3 O plasmídeo construído de *E. coli* pBR322. Observe a localização de alguns importantes sítios de restrição – para *Pst*I, *Eco*RI, *Bam*HI, *Sal*I e *Pvu*II; genes com resistência para ampicilina e tetraciclina; e a origem de replicação (ori). Construído em 1977, este foi um dos plasmídeos expressamente projetados para clonagem em *E. coli*.

As origens de replicação inseridas em vetores de plasmídeos comuns foram originalmente derivadas de plasmídeos que ocorrem naturalmente. Como em pBR322, cada uma dessas origens é regulada para manter um número de cópias do plasmídeo específico. Dependendo da origem utilizada, o número de cópias do plasmídeo pode variar de uma a centenas ou milhares por célula, proporcionando muitas opções para os investigadores. Dois plasmídeos diferentes não podem funcionar na mesma célula utilizando a mesma origem de replicação, porque a regulação de um irá interferir com a replicação do outro. Esses plasmídeos são considerados incompatíveis. Quando um pesquisador deseja introduzir dois ou mais plasmídeos diferentes em uma célula bacteriana, cada plasmídeo deve ter uma origem de replicação diferente.

No laboratório, os pequenos plasmídeos podem ser introduzidos em células bacterianas em um processo chamado de **transformação**. As células (frequentemente *E. coli,* mas outras espécies bacterianas também são utilizadas) e o DNA plasmidial são incubados juntos a 0°C em uma solução de cloreto de cálcio e, em seguida, submetidos a choque térmico, alterando rapidamente a temperatura para entre 37°C e 43°C. Por motivos não muito bem compreendidos, algumas das células tratadas desse modo captam o DNA plasmidial. Algumas espécies de bactérias, como o *Acinetobacter baylyi*, são naturalmente competentes na captação do DNA e não necessitam de tratamento com cloreto de cálcio – choque térmico. Em um método alternativo, as células incubadas com o DNA plasmidial são submetidas a um pulso elétrico de alta voltagem. Esse método, chamado de **eletroporação**, torna a membrana bacteriana temporariamente permeável a moléculas grandes.

Independente da abordagem, relativamente poucas células captam o DNA plasmidial, de modo que um método é necessário para identificar aquelas que o fazem. A estratégia é utilizar um dos dois tipos de genes no plasmídeo, chamados de marcadores de seleção e triagem. Os **marcadores de seleção** podem permitir tanto o crescimento de uma célula (seleção positiva) quanto matar a célula (seleção negativa), sob um conjunto definido de condições. O plasmídeo pBR322 fornece exemplos das seleções positiva e negativa (**Figura 9-4**). Um **marcador de triagem** é um gene que codifica uma proteína que leva a célula a produzir uma molécula colorida ou fluorescente. As células não são prejudicadas quando o gene está presente, e as células que transportam o plasmídeo são facilmente identificadas pelas colônias coloridas ou fluorescentes que produzem.

A transformação de células bacterianas normais com DNA purificado (processo pouco eficaz) torna-se menos bem-sucedida com o aumento do tamanho do plasmídeo, sendo difícil clonar segmentos de DNA maiores do que cerca de 15.000 pb, quando os plasmídeos são utilizados como o vetor.

Para ilustrar a utilização de um plasmídeo como vetor de clonagem, considere um gene bacteriano típico, que codifica uma recombinase chamada de proteína RecA (ver Capítulo 25). Na maior parte das bactérias, o gene que codifica RecA é um dos milhares de outros genes em um cromos-

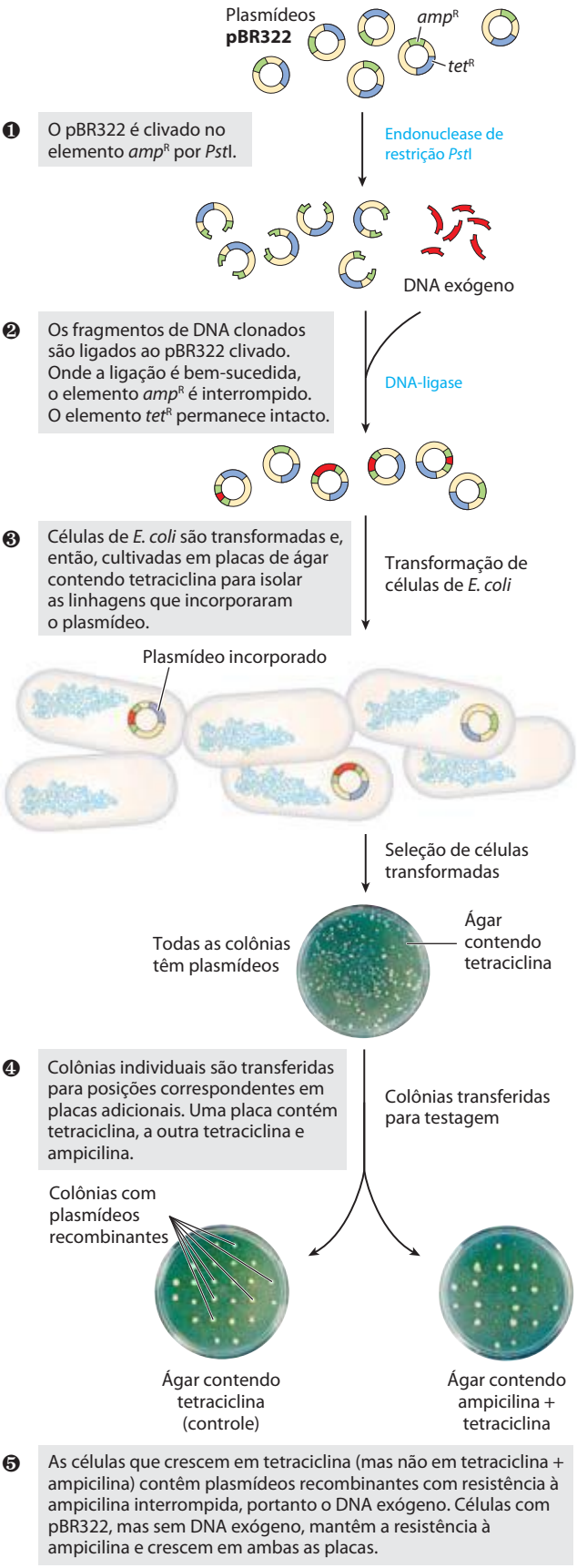


FIGURA 9-4 Utilização do pBR322 para clonar DNA exógeno em *E. coli* e identificar as células que o contêm. Clonagem de plasmídeo

somo com milhares de pares de bases de comprimento. O gene *recA* tem um pouco mais de 1.000 pb de comprimento. Um plasmídeo seria uma boa escolha para a clonagem de um gene desse tamanho. Como descrito posteriormente, o gene clonado pode ser modificado de vários modos, e as variantes do gene podem ser expressas em níveis elevados para permitir a purificação da proteína codificada.

Cromossomos artificiais bacterianos Projetos de sequenciamento de grandes genomas frequentemente exigem a clonagem de segmentos de DNA mais longos do que normalmente podem ser incorporados em vetores de clonagem plasmidiais padrão, tais como o pBR322. Para atender a essa necessidade, foram desenvolvidos vetores plasmidiais com características especiais que permitem a clonagem de segmentos muito longos (normalmente de 100.000 a 300.000 pb) de DNA. Quando esses grandes segmentos de DNA clonado são adicionados, esses vetores são suficientemente grandes para serem considerados cromossomos e são chamados de **cromossomos artificiais bacterianos**, ou **BAC** (**Figura 9-5**).

Um vetor BAC (sem qualquer DNA clonado inserido) é um plasmídeo relativamente simples, em geral não muito maior do que outros vetores plasmidiais. Para acomodar segmentos de DNA clonado muito longos, os vetores BAC têm origens estáveis de replicação que mantêm o plasmídeo em uma ou duas cópias por célula. O baixo número de cópias é útil para a clonagem de grandes segmentos de DNA porque limita as oportunidades de reações de recombinação indesejadas que podem de modo imprevisto alterar grandes DNAs clonados ao longo do tempo. Os BACs também incluem genes *par* que codificam proteínas, as quais direcionam a distribuição confiável de cromossomos recombinantes para as células-filhas na divisão celular, aumentando assim a probabilidade de cada uma delas transportar uma cópia, mesmo quando algumas cópias estão presentes. O vetor BAC inclui tanto os marcadores de seleção quanto os marcadores de triagem. O vetor BAC mostrado na Figura 9-5 contém um gene que confere resistência ao antibiótico cloranfenicol (Cm^R). A seleção positiva para vetores contendo células ocorre em placas de ágar contendo esse antibiótico. Um gene *lacZ*, necessário para a produção da enzima β-galactosidase, é um marcador de triagem capaz de revelar as células que contêm os plasmídeos – agora cromossomos – que incorporam os segmentos de DNA clonados. A β-galactosidase catalisa a conversão da molécula incolor 5-bromo-4-cloro-3-indol-β-D-galactopiranosídeo (X-gal) a um produto azul. Se o gene intacto for expresso, a colônia que o contém será azul. Se a expressão do gene é interrompida pela introdução de um segmento de DNA clonado, a colônia será branca.

Cromossomos artificiais de levedura Assim como no caso de *E. coli*, a genética de leveduras é uma área de conhecimento bem desenvolvida. O genoma de *Saccharomyces cerevisiae* contém apenas 14×10^6 pb (menos que quatro vezes o tamanho do cromossomo de *E. coli*), e sua sequência inteira é conhecida. A levedura também é muito fácil de ser mantida e cultivada em larga escala em um laboratório. Vetores plasmidiais têm sido construídos para leveduras, empregando os mesmos princípios que se aplicam para o uso de vetores de *E. coli*. Métodos convenientes para mover o

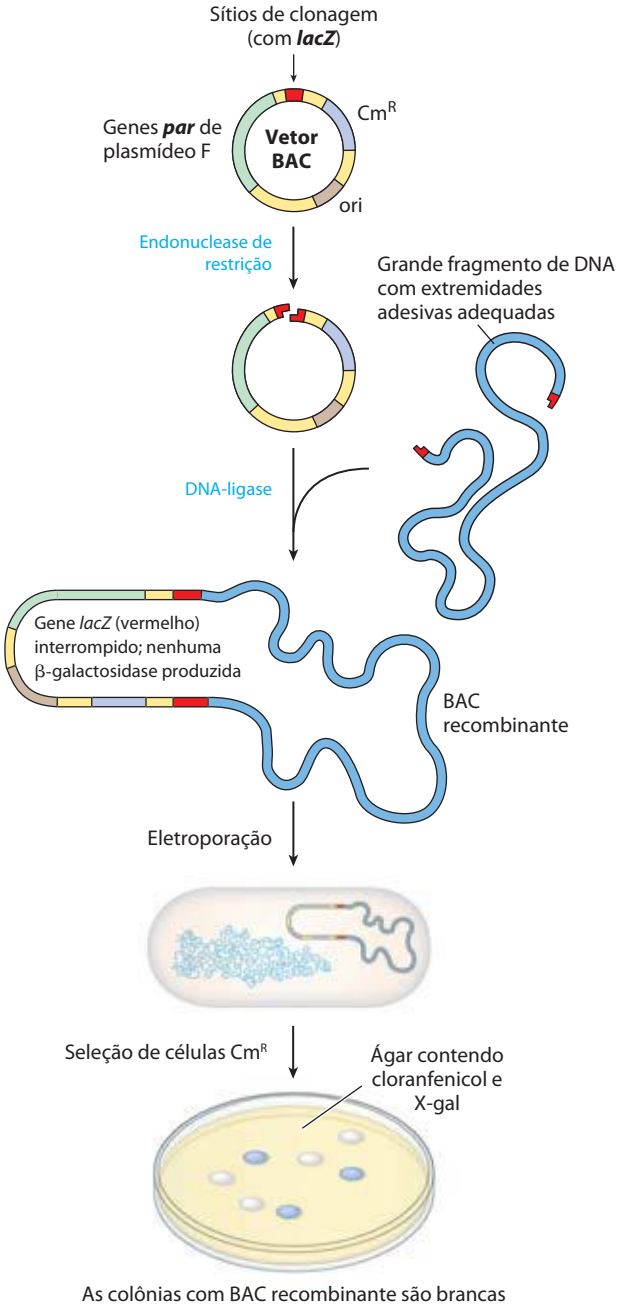


FIGURA 9-5 Cromossomos artificiais bacterianos (BAC) como vetores de clonagem. O vetor é um plasmídeo relativamente simples, com uma origem de replicação (ori) que direciona a replicação. Os genes *par,* derivados de um tipo de plasmídeo chamado de plasmídeo F, auxiliam na distribuição dos plasmídeos para as células-filhas, na divisão celular. Isso aumenta a probabilidade de cada célula-filha carregar uma cópia do plasmídeo, mesmo quando poucas cópias estão presentes. O baixo número de cópias é útil na clonagem de grandes segmentos de DNA porque limita as oportunidades para que reações de recombinação indesejadas possam alterar de modo imprevisível grandes DNAs clonados ao longo do tempo. O BAC inclui marcadores de seleção. Um gene *lacZ* (necessário à produção da enzima β-galactosidase) se situa na região de clonagem de tal modo que ele é inativado pelas inserções de DNA clonado. A introdução de BACs recombinantes nas células por eletroporação é promovida pelo de uso de células com uma parede celular alterada (mais porosa). Os DNAs recombinantes são triados para resistência ao antibiótico cloranfenicol (Cm^R). As placas também contêm X-gal, um substrato para β-galactosidase que gera um produto azul. As colônias com β-galactosidase ativas e, portanto, nenhum DNA inserido no vetor BAC se tornam azuis; as colônias sem atividade de β-galactosidase e, portanto, com o DNA desejado inserido são brancas.

DNA para dentro e para fora das células de levedura permitem o estudo de muitos aspectos da bioquímica de células eucarióticas. Alguns plasmídeos recombinantes incorporam múltiplas origens de replicação e outros elementos que lhes permitem ser utilizados em mais de uma espécie (p. ex., em leveduras e em *E. coli*). Esses plasmídeos que podem ser propagados em células de duas ou mais espécies são chamados de **vetores *shuttle***.

As pesquisas em grandes genomas e a necessidade associada de vetores de clonagem de alta capacidade levaram ao desenvolvimento de **cromossomos artificiais de levedura** ou **YAC** (**Figura 9-6**). Os vetores YAC contêm todos os elementos necessários para manter um cromossomo eucariótico no núcleo da levedura: uma origem de replicação de levedura, dois marcadores de seleção e sequências especializadas (derivadas de telômeros e do centrômero) necessárias para a estabilidade e a segregação adequadas dos cromossomos na divisão celular (ver Capítulo 24). Na preparação para seu uso na clonagem, o vetor é propagado como um plasmídeo bacteriano circular e, então, isolado e purificado. A clivagem com endonuclease de restrição (*Bam*HI na Figura 9-6) remove um segmento do DNA entre duas sequências de telômeros (TEL), deixando os telômeros nas extremidades do DNA linearizado. A clivagem em outro sítio interno (por *Eco*RI na Figura 9-6) divide o vetor em dois segmentos de DNA, chamados de braços do vetor, cada qual com um marcador de seleção diferente.

O DNA genômico a ser clonado é preparado por digestão parcial com endonucleases de restrição para obter um tamanho de fragmento adequado. Em seguida, os fragmentos genômicos são separados por **eletroforese em gel de campo pulsado**, uma variação da eletroforese em gel (ver Figura 3-18) que separa os segmentos de DNA muito grandes. Os fragmentos de DNA de tamanho adequado (até cerca de 2×10^6 pb) são misturados com os braços do vetor preparado e ligados. A mistura de ligação é então utilizada para transformar células (pré-tratadas para degradar parcialmente suas paredes celulares) com essas moléculas de DNA muito grandes – que agora têm estrutura e tamanho para serem consideradas cromossomos de levedura. A cultura em um meio que exige a presença de ambos os genes marcadores de seleção assegura o crescimento somente das células de levedura que contêm um cromossomo artificial com grande inserção colocada entre os dois braços do vetor (Figura 9-6). A estabilidade dos clones YAC aumenta com o comprimento do segmento de DNA clonado (até certo ponto). Aqueles contendo inserções com mais de 150.000 pb são quase tão estáveis quanto cromossomos celulares normais, ao passo que os com inserções de menos de 100.000 pb de comprimento são gradativamente perdidos durante a mitose (portanto, em geral não existe nenhum clone de células de levedura carregando apenas as extremidades de dois vetores liga-

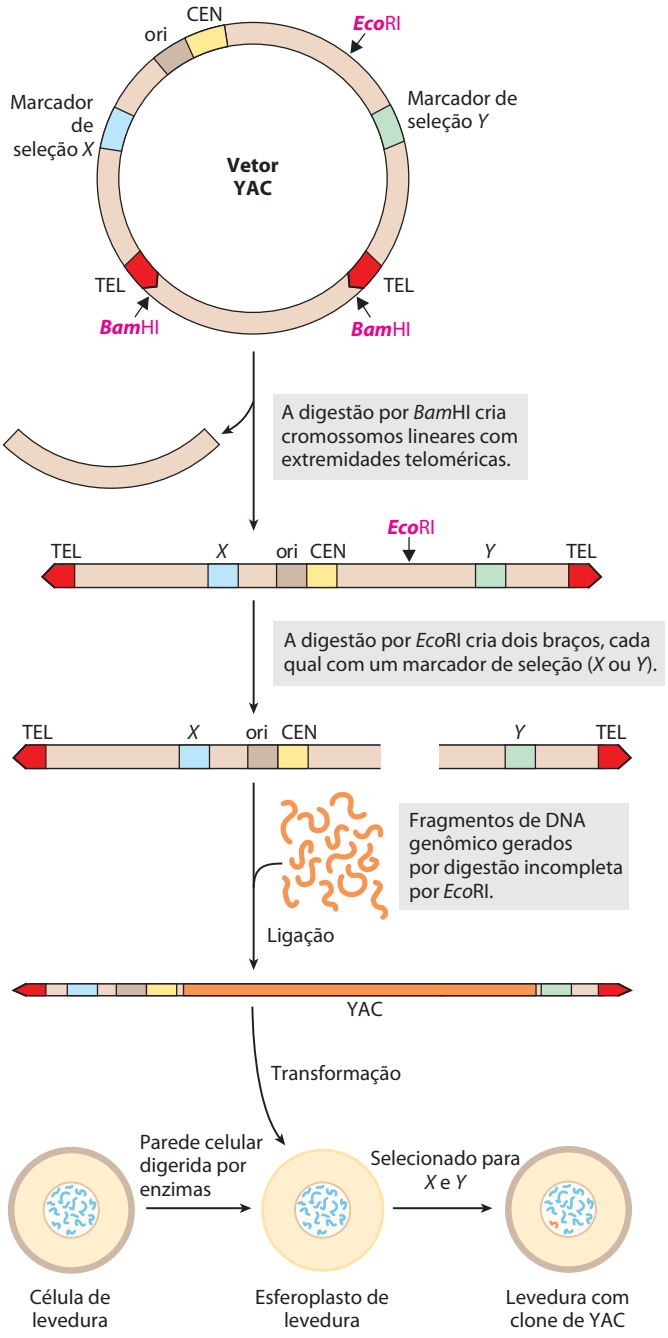


FIGURA 9-6 Construção de um cromossomo artificial de levedura (YAC). Um vetor YAC inclui uma origem de replicação (ori), um centrômero (CEN), dois telômeros (TEL) e marcadores de seleção (*X* e *Y*). A digestão por *Bam*HI e *Eco*RI gera dois braços de DNA separados, cada um com uma extremidade telomérica e um marcador de seleção. Um grande segmento de DNA (p. ex., até 2×10^6 pb de um genoma humano) se liga aos dois braços para criar um cromossomo artificial de levedura. O YAC transforma as células de levedura (preparadas pela remoção da parede celular para formar esferoplastos) e as células são selecionadas para *X* e *Y*; as células sobreviventes propagam as inserções de DNA.

dos juntos ou vetores com apenas inserções curtas). Os YAC sem um telômero em cada uma das extremidades se degradam rapidamente.

Como acontece com os BACs, os vetores YAC podem ser utilizados para clonar segmentos de DNA muito longos. Além disso, o DNA clonado em um YAC pode ser modificado para o estudo de funções de sequências especializadas no metabolismo de cromossomos, mecanismos de regulação e expressão gênica e muitos outros problemas na biologia molecular eucariótica.

Genes clonados podem ser expressos para amplificar a produção de proteínas

Frequentemente o produto de um gene clonado, mais que o próprio gene, é o interesse principal, sobretudo quando a proteína tem valor comercial, terapêutico ou de pesquisa. Os bioquímicos utilizam proteínas purificadas para muitos fins, incluindo a descoberta do seu funcionamento, o estudo de mecanismos de reação, a geração de anticorpos para as proteínas, a reconstituição de atividades celulares complexas no tubo de ensaio com componentes purificados e o exame de parceiros de ligação de proteínas. Com uma compreensão maior sobre os fundamentos do DNA, do RNA e do metabolismo proteico e sua regulação em um organismo hospedeiro como a *E. coli* ou leveduras, os pesquisadores manipulam células para expressar genes clonados a fim de estudar seus produtos proteicos. O objetivo geral é modificar as sequências ao redor de um gene clonado para enganar o organismo hospedeiro a fim de produzir o produto proteico do gene, frequentemente em níveis muito elevados. A superexpressão de uma proteína torna sua purificação subsequente muito mais fácil.

Como exemplo, será utilizada a expressão de uma proteína eucariótica em uma bactéria. Os genes eucarióticos têm sequências circundantes necessárias para sua transcrição e regulação nas células dos quais são derivados, mas essas sequências não funcionam em bactérias. Assim, os genes eucarióticos não têm os elementos de sequência de DNA necessários para sua expressão controlada em células bacterianas – promotores (sequências que informam à RNA-polimerase onde se ligar para iniciar a síntese de mRNA), sítios de ligação de ribossomos (sequências que permitem a tradução de mRNA para proteína) e sequências regulatórias adicionais. Portanto, as sequências regulatórias bacterianas adequadas para a transcrição e tradução devem ser inseridas no DNA vetor nas posições corretas em relação ao gene eucariótico. Em alguns casos, genes clonados são expressos de maneira tão eficiente que seus produtos proteicos representam 10% ou mais da proteína celular. Nessas concentrações, algumas proteínas estranhas podem matar a célula hospedeira (geralmente *E. coli*); portanto, a expressão do gene clonado deve ser limitada a poucas horas antes da coleta das células.

Vetores de clonagem com sinais de transcrição e tradução necessários à expressão regulada de um gene clonado são chamados de **vetores de expressão**. A taxa de expressão do gene clonado é controlada pela substituição do promotor normal e sequências regulatórias do gene por versões mais eficientes e convenientes fornecidas pelo vetor. Geralmente, um promotor bem caracterizado e seus elementos regulatórios são posicionados próximos a vários sítios de restrição exclusivos para a clonagem, de modo que

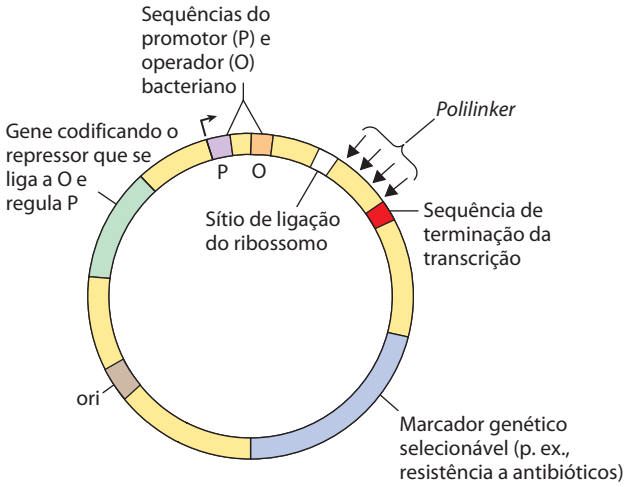


FIGURA 9-7 Sequências de DNA em um típico vetor de expressão de *E. coli*. O gene a ser expresso é inserido em um dos sítios de restrição no *polilinker*, perto do promotor (P), com a extremidade do gene codificando a terminação amina da proteína posicionada o mais perto possível do promotor. O promotor permite a transcrição eficiente do gene inserido, e a sequência de transcrição-terminação algumas vezes aumenta a quantidade e a estabilidade do mRNA produzido. O operador (O) permite a regulação por um repressor que se liga a ele. O sítio de ligação do ribossomo fornece sinais de sequências para a tradução eficiente do mRNA derivado do gene. O marcador selecionado permite a seleção de células contendo o DNA recombinante.

genes inseridos nos sítios de restrição sejam expressos a partir desses elementos do promotor regulado (**Figura 9-7**). Alguns desses vetores incorporam outras características, tais como um sítio de ligação de ribossomos bacterianos, para potencializar a tradução do mRNA derivado do gene (Capítulo 27) ou uma sequência de terminação da transcrição (Capítulo 26).

Muitos sistemas diferentes são utilizados para expressar proteínas recombinantes

Todo organismo vivo tem a capacidade de expressar genes em seu DNA genômico; portanto, em princípio, qualquer organismo pode ser um hospedeiro para expressar proteínas de espécies diferentes (heterólogas). De fato, quase todos os tipos de organismos são utilizados para esse fim, e cada tipo de hospedeiro tem um conjunto próprio de vantagens e desvantagens.

Bactérias As bactérias, especialmente *E. coli*, são os hospedeiros mais comuns para a expressão de proteínas. As sequências regulatórias que determinam a expressão gênica em *E. coli* e em muitas outras bactérias são bem compreendidas e podem ser aproveitadas para expressar proteínas clonadas em níveis elevados. As bactérias são fáceis de armazenar e cultivar em laboratório, em meios de cultivo baratos. Também existem métodos eficientes para colocar DNA em bactérias e extrair o DNA delas. As bactérias podem ser cultivadas em grandes quantidades em fermentadores comerciais, fornecendo uma rica fonte de proteínas clonadas. Entretanto, existem problemas. Quando expressas em bactérias, algumas proteínas heterólogas não se dobram corretamente, e muitas não fazem as ligações covalentes ou a clivagem proteolítica necessárias à sua atividade. Devido a certas características de uma sequência gênica, dificilmente um gene específico também se expressa em bactérias. Por exemplo, regiões intrinsecamente desordenadas são mais comuns em proteínas eucarióticas. Expressas em bactérias, muitas proteínas eucarióticas se agregam em precipitados celulares insolúveis chamados de corpos de inclusão. Por essas e muitas outras razões, algumas proteínas eucarióticas são inativas quando purificadas de bactérias ou não se expressam de modo algum. Para ajudar a lidar com alguns desses problemas, novas linhagens bacterianas hospedeiras são regularmente desenvolvidas para incluir melhorias, como a presença de chaperonas proteicas eucarióticas ou enzimas que modifiquem as proteínas eucarióticas.

Há muitos sistemas especializados para expressar proteínas em bactérias. O promotor e as sequências regulatórias associadas ao óperon lactose (ver Capítulo 28) são com frequência fusionados ao gene de interesse para direcionar a transcrição. O gene clonado será transcrito quando a lactose for adicionada ao meio de cultivo. Entretanto, a regulação no sistema lactose é "frouxa"; ela não é desligada completamente quando a lactose está ausente – problema potencial se o produto do gene clonado for tóxico para as células hospedeiras. A transcrição a partir do promotor Lac também não é muito eficiente para algumas aplicações.

Um sistema alternativo utiliza um promotor e a RNA-polimerase encontrada em um vírus bacteriano chamado de bacteriófago T7. Se o gene clonado for fusionado a um promotor T7, ele não será transcrito pela RNA-polimerase de *E. coli,* mas pela RNA-polimerase T7. O gene que codifica essa polimerase é clonado separadamente na mesma célula em uma construção que proporciona uma regulação rigorosa (permitindo a produção controlada da RNA-polimerase

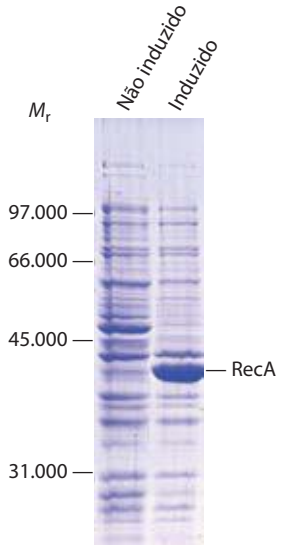


FIGURA 9-8 Expressão regulada da proteína RecA em uma célula bacteriana. O gene que codifica a proteína RecA, fusionado com o promotor T7 do bacteriófago, é clonado em um vetor de expressão. Em condições normais de crescimento (não induzido), nenhuma proteína RecA aparece. Quando a RNA-polimerase de T7 é induzida na célula, o gene *recA* é expresso e grandes quantidades da proteína RecA são produzidas. As posições dos marcadores de peso molecular padrão que correm no mesmo gel são indicadas.

T7). A polimerase também é muito eficiente e direciona níveis elevados de expressão da maioria dos genes fusionados no promotor T7. Esse sistema é utilizado para expressar a proteína RecA em células bacterianas (**Figura 9-8**).

Leveduras A levedura *Saccharomyces cerevisiae* é provavelmente o organismo eucariótico mais bem compreendido e um dos mais fáceis de cultivar e manipular em laboratório. Como as bactérias, essa levedura pode crescer em meios simples e baratos. A levedura tem paredes celulares resistentes, que são difíceis de romper para introduzir vetores de DNA; assim, as bactérias são mais convenientes para fazer a maior parte da engenharia genética e manutenção de vetores. É por isso que o vetor de leveduras foi propagado pela primeira vez em bactérias. Existem vários excelentes vetores de transporte para esse fim.

Os princípios subjacentes à expressão de uma proteína em leveduras são os mesmos daqueles em bactérias. Os genes clonados devem ser ligados aos promotores que podem direcionar a expressão de alto nível em leveduras. Por exemplo, os genes da levedura *GAL1* e *GAL10* são regulados pela célula de modo a serem expressos quando as células da levedura são cultivadas em meios com galactose, mas desligados quando as células são cultivadas em glicose. Assim, se um gene heterólogo é expresso utilizando as mesmas sequências reguladoras, a expressão desse gene pode ser controlada simplesmente pela escolha de um meio adequado para o crescimento celular.

Alguns dos mesmos problemas que acompanham a expressão de proteínas em bactérias também ocorrem em leveduras. Proteínas heterólogas podem não se dobrar corretamente, a levedura talvez não tenha as enzimas necessárias para modificar as proteínas em suas formas ativas, ou a expressão de proteínas pode ser dificultada por algumas características da sequência gênica. Entretanto, como o *S. cerevisiae* é um eucarioto, a expressão de genes eucarióticos (especialmente os genes de leveduras) é algumas vezes mais eficiente nesse hospedeiro do que em bactérias. O enovelamento e a modificação dos produtos também podem ser mais precisos do que nas proteínas expressas em bactérias.

Insetos e vírus de insetos Os baculovírus são vírus de insetos com genomas de DNA de cadeia dupla. Ao infectarem larvas de insetos hospedeiras, agem como parasitas, matando as larvas e transformando-as em fábricas de produção de vírus. No final do processo de infecção, os vírus produzem grandes quantidades de duas proteínas (p10 e poliedrina), nenhuma das quais é necessária para a produção do vírus em células de inseto cultivadas. Os genes para ambas as proteínas podem ser substituídos pelo gene de uma proteína heteróloga. Quando o vírus recombinante resultante é utilizado para infectar células ou larvas de insetos, a proteína heteróloga é frequentemente produzida em níveis muito elevados – até 25% do total de proteínas presentes no final do ciclo de infecção.

O nucleopoliedrovírus multicapsídeo *Autographa californica* (AcMNPV) é o **baculovírus** mais utilizado para a expressão de proteínas. Ele tem um genoma grande demais (134.000 pb) para a clonagem direta. A purificação de vírus também é complicada. Esses problemas foram resolvidos com a criação de **bacmídeos**, DNA circulares grandes que incluem todo o genoma do baculovírus, além das sequências que permitem a replicação do bacmídeo em *E. coli* (**Figura 9-9**). O gene de interesse é clonado em um plasmídeo menor e combinado com o plasmídeo maior por recombinação sítio-específica *in vivo* (ver Figura 25-37). O bacmídeo recombinante é então isolado e transfectado para as células de insetos (o termo **transfecção** é utilizado quando o DNA utilizado para a transformação inclui sequências virais e leva à replicação viral), seguido por uma recuperação da proteína assim que termina o ciclo de infecção. Uma grande variedade de sistemas de bacmídeos está disponível comercialmente. Os sistemas de baculovírus não são bem-sucedidos com todas as proteínas. Entretanto, com esses sistemas, as células de insetos algumas vezes replicam com sucesso os padrões de modificação de proteína de eucariotos superiores e produzem proteínas eucarióticas ativas corretamente modificadas.

Células de mamíferos em cultura O modo mais conveniente para se introduzir genes clonados em uma célula de mamífero é por meio de vírus. Esse método tem a vantagem de utilizar a capacidade natural de um vírus para inserir seu DNA ou RNA em uma célula e, por vezes, no cromossomo celular. Diversos vírus de mamíferos geneticamente modificados estão disponíveis como vetores, incluindo o adenovírus e retrovírus humanos. O gene de interesse é clonado, de modo que a sua expressão seja controlada por um promotor do vírus. O vírus utiliza seus mecanismos de infecção naturais para introduzir o genoma recombinante nas células, onde se expressa a proteína clonada. Esses sistemas têm a vantagem de as proteínas serem expressas transitoriamente (se o DNA viral for mantido separadamente a partir do genoma da célula hospedeira e, por fim, degradado) ou permanentemente (se o DNA viral for integrado no genoma da célula hospedeira). Com a escolha correta da célula hospedeira, a modificação pós-tradução adequada da proteína para sua forma ativa pode ser assegurada. Entretanto, o crescimento de células de mamíferos em culturas de tecidos é muito caro, e essa tecnologia é geralmente utilizada para testar a função de uma proteína *in vivo*, em vez de produzir uma proteína em grandes quantidades.

A alteração de genes clonados produz proteínas alteradas

Técnicas de clonagem podem ser utilizadas não só para produzir mais proteínas, mas para produzir produtos proteicos alterados, sutilmente ou de forma drástica, a partir de suas formas nativas. Aminoácidos específicos podem ser substituídos individualmente por **mutagênese sítio-direcionada**. Essa técnica melhorou substancialmente as pesquisas sobre proteínas, ao permitir que os investigadores façam alterações específicas na estrutura primária e analisem os efeitos dessas alterações no enovelamento, na estrutura tridimensional e na atividade de proteínas. Essa abordagem poderosa para estudar a estrutura e função das proteínas muda a sequência de aminoácidos pela alteração da sequência de DNA do gene clonado. Se os sítios de restrição adequados se situam ao lado da sequência a ser alterada, os pesquisadores podem simplesmente remover um segmento de DNA e substituí-lo por um sintético idêntico ao original, exceto para a mudança desejada (**Figura 9-10a**).

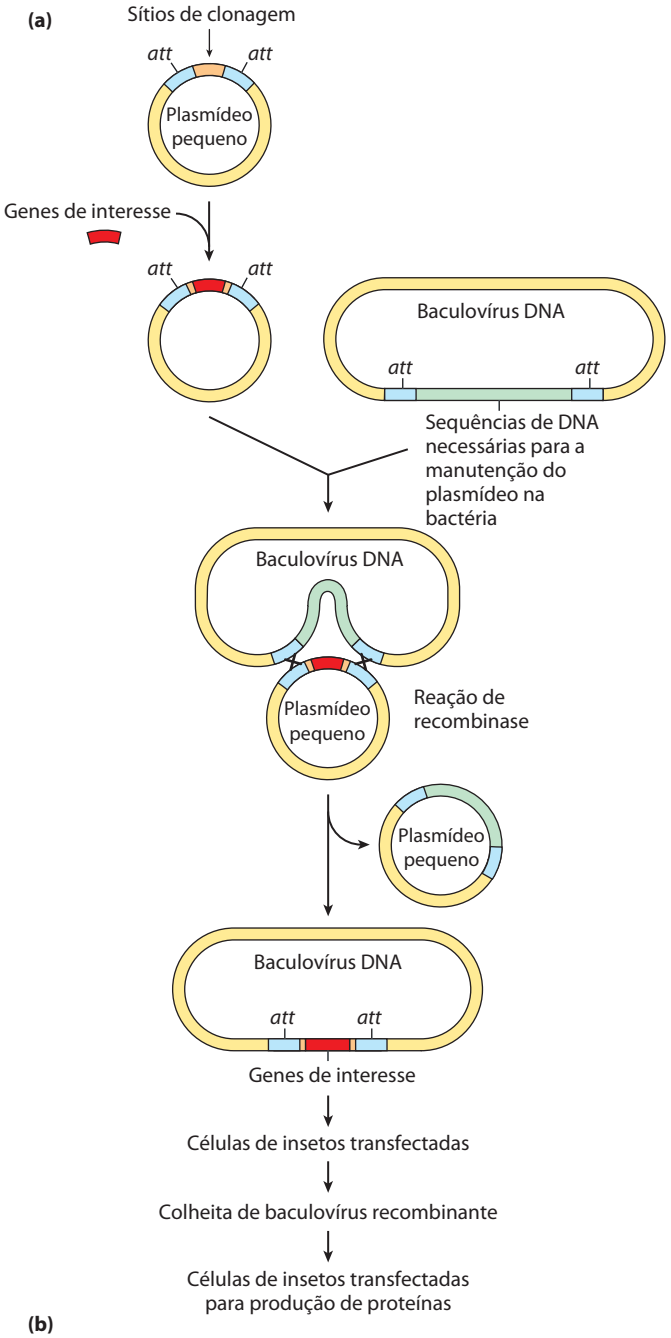


FIGURA 9-9 **Clonagem com baculovírus.** **(a)** Aqui é mostrada a construção de um vetor típico utilizado para a expressão de proteínas em baculovírus. O gene de interesse é clonado em um plasmídeo pequeno (esquerda) entre dois sítios (*att*) reconhecidos pela recombinase sítio-específica, sendo em seguida introduzido no vetor do baculovírus por recombinação sítio-específica (ver Figura 25-37). Isso gera um produto de DNA circular utilizado para infectar as células de uma larva de inseto. O gene de interesse é expresso durante o ciclo de infecção, abaixo de um promotor que normalmente expressa uma proteína de revestimento de baculovírus em níveis muito elevados. **(b)** As fotografias mostram (à esquerda) uma larva de inseto infectada com o vetor do baculovírus recombinante expressando uma proteína que produz cor vermelha e (à direita) uma larva não infectada.

Na ausência de sítios de restrição adequadamente localizados, a **mutagênese oligonucleotídeo-direcionada** pode criar uma alteração específica na sequência de DNA (Figura 9-10b). Duas cadeias de DNA curtas, sintetizadas de modo complementar, cada uma com a alteração de base desejada, são renaturadas em cadeias opostas do gene clonado dentro de um vetor de DNA circular adequado. O malpareamento de um único par de bases em 30 a 40 pb não impede a renaturação. Os dois oligonucleotídeos renaturados sintetizam o primeiro DNA em ambas as direções ao redor do vetor de plasmídeo, criando duas fitas complementares que contêm a mutação. Após vários ciclos de amplificação seletiva, utilizando uma técnica chamada de reação em cadeia da polimerase (PCR; descrita na p. 327), a mutação contendo o DNA predomina na população e pode ser utilizada para transformar bactérias. A maior parte das bactérias transformadas terá plasmídeos portadores da mutação. Se necessário, o molde do DNA do plasmídeo não mutante pode ser seletivamente eliminado por clivagem com a enzima de restrição *Dpnl*. O plasmídeo-molde, geralmente isolado a partir de um tipo selvagem de *E. coli*, tem um resíduo A metilado em cada cópia do palíndromo de quatro nucleótidos, GATC (chamado de sítio dam; ver Figura 25-21). O novo DNA contendo a mutação não tem resíduos A metilados porque a replicação é realizada *in vitro*. A *Dpnl* cliva seletivamente o DNA na sequência GATC apenas se o resíduo A de uma ou ambas as cadeias for metilado, isto é, ela degrada apenas o molde.

No caso do gene *rec*A bacteriano, o produto desse gene, a proteína RecA, tem várias atividades (ver Seção 25.3). Ela liga e forma uma estrutura filamentosa no DNA, alinha dois DNAs de sequências semelhantes e hidrolisa o ATP. Um resíduo de aminoácido especial na RecA (polipeptídeo de 352 resíduos), o resíduo Lys na posição 72, está envolvido na hidrólise do ATP. Alterando o Lys^{72} para uma Arg, uma variante da proteína RecA é criada e irá ligar, mas não hidrolisar o ATP (Figura 9-10c). A engenharia e a purificação dessa proteína variante RecA tem facilitado a investigação sobre os papéis da hidrólise do ATP no funcionamento dessa proteína.

Em genes podem ser introduzidas alterações que envolvem mais de um par de bases. Grandes partes de um gene podem ser excluídas cortando-se um segmento com endonucleases de restrição e ligando as porções remanescentes para formar um gene menor. Por exemplo, se uma proteína tiver dois domínios, o segmento do gene que codifica um dos domínios pode ser removido para produzir uma proteína que agora tem apenas um dos dois originais.

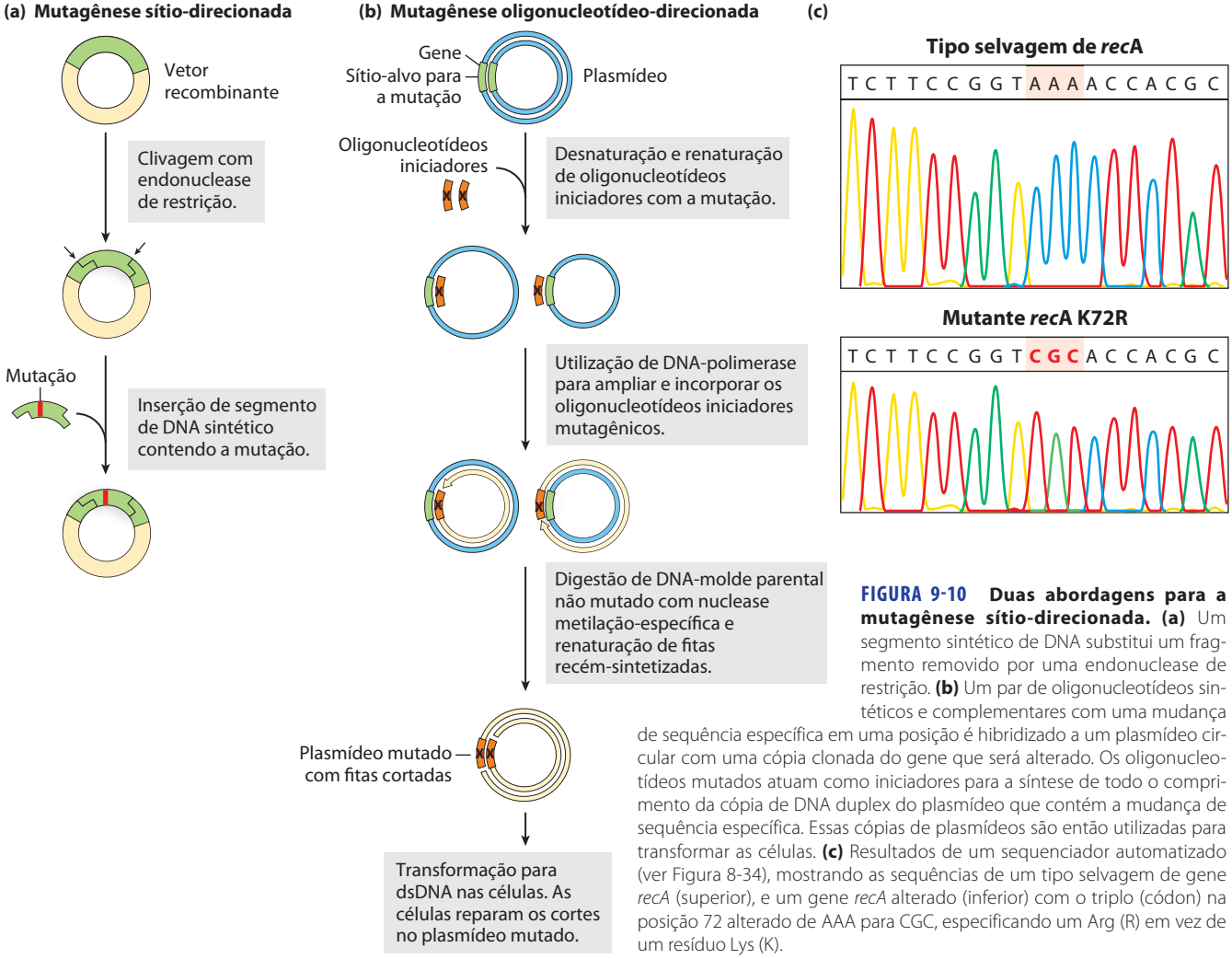


FIGURA 9-10 Duas abordagens para a mutagênese sítio-direcionada. (a) Um segmento sintético de DNA substitui um fragmento removido por uma endonuclease de restrição. **(b)** Um par de oligonucleotídeos sintéticos e complementares com uma mudança de sequência específica em uma posição é hibridizado a um plasmídeo circular com uma cópia clonada do gene que será alterado. Os oligonucleotídeos mutados atuam como iniciadores para a síntese de todo o comprimento da cópia de DNA duplex do plasmídeo que contém a mudança de sequência específica. Essas cópias de plasmídeos são então utilizadas para transformar as células. **(c)** Resultados de um sequenciador automatizado (ver Figura 8-34), mostrando as sequências de um tipo selvagem de gene *recA* (superior), e um gene *recA* alterado (inferior) com o triplo (códon) na posição 72 alterado de AAA para CGC, especificando um Arg (R) em vez de um resíduo Lys (K).

Partes dos dois genes diferentes podem ser ligadas para criar novas combinações; o produto desse gene fusionado é chamado de **proteína de fusão**. Pesquisadores têm métodos engenhosos para realizar praticamente qualquer alteração genética *in vitro*. Após a reintrodução do DNA alterado na célula, eles investigam as consequências dessa alteração.

Marcadores terminais fornecem instrumentos para a purificação por afinidade

A cromatografia de afinidade é um dos métodos mais eficientes para a purificação de proteínas (ver Figura 3-17c). Infelizmente, muitas proteínas não se ligam a um ligante que possa ser convenientemente imobilizado em uma matriz de coluna. Entretanto, o gene para quase todas as proteínas pode ser alterado para expressar uma proteína de fusão purificada por cromatografia de afinidade. O gene que codifica a proteína-alvo é fusionado com um gene que codifica um peptídeo ou proteína que se liga a um ligante simples e estável, com elevada afinidade e especificidade. O peptídeo ou proteína utilizado para essa finalidade é denominado **marcador**. As sequências dos marcadores são adicionadas a genes de tal modo que as proteínas resultantes tenham marcadores em seu grupo amino ou carboxila terminal. A Tabela 9-3 lista alguns dos peptídeos ou proteínas comumente utilizados como marcadores.

TABELA 9-3 Caudas proteicas comumente usadas

Marcadores proteicos/ peptídicos	Massa molecular (kDa)	Ligante imobilizado
Proteína A	59	Porção Fc da IgG
$(His)_6$	0,8	Ni^{2+}
Glutationa-*S*-transferase (GST)	26	Glutationa
Proteína ligante de maltose	41	Maltose
β-Galactosidase	116	*p*-Aminofenil-β-D-tiogalactosídeo (TPEG)
Domínio ligante de quitina	5,7	Quitina

O procedimento geral pode ser ilustrado focalizando um sistema que utiliza a glutationa-S-transferase (GST) (**Figura 9-11**). A GST é uma pequena enzima (M_r 26.000) que se liga fortemente e de modo específico à glutationa. Quando a sequência do gene de GST é fusionada ao gene-alvo, a proteína de fusão adquire a capacidade de se ligar à glutationa. A proteína de fusão é expressa em um organismo hospedeiro, tal como uma bactéria, e um extrato bruto é preparado. Uma coluna é preenchida com matriz porosa, que consiste em um ligante (glutationa) imobilizado por esferas microscópicas de um polímero reticulado estável, como a agarose. À medida que o extrato bruto se infiltra por essa matriz, a proteína de fusão é imobilizada por sua ligação à glutationa. As outras proteínas no extrato são lavadas e descartadas da coluna. A interação entre GST e glutationa é firme, porém não covalente, permitindo que a proteína de fusão seja eluída suavemente da coluna com uma solução que contém alta concentração de sais ou de glutationa livre para competir com o ligante imobilizado pela ligação com a GST. A proteína de fusão é obtida muitas vezes com boa produção e alta pureza. Em alguns sistemas disponíveis comercialmente, o marcador pode ser removido em grande parte ou totalmente da proteína de fusão purificada por uma protease que cliva uma sequência próxima da junção entre a proteína-alvo e seu marcador.

Um marcador mais curto, com ampla aplicação, consiste em uma sequência simples de seis ou mais resíduos His. Esses marcadores de histidina, ou marcadores His, se ligam de

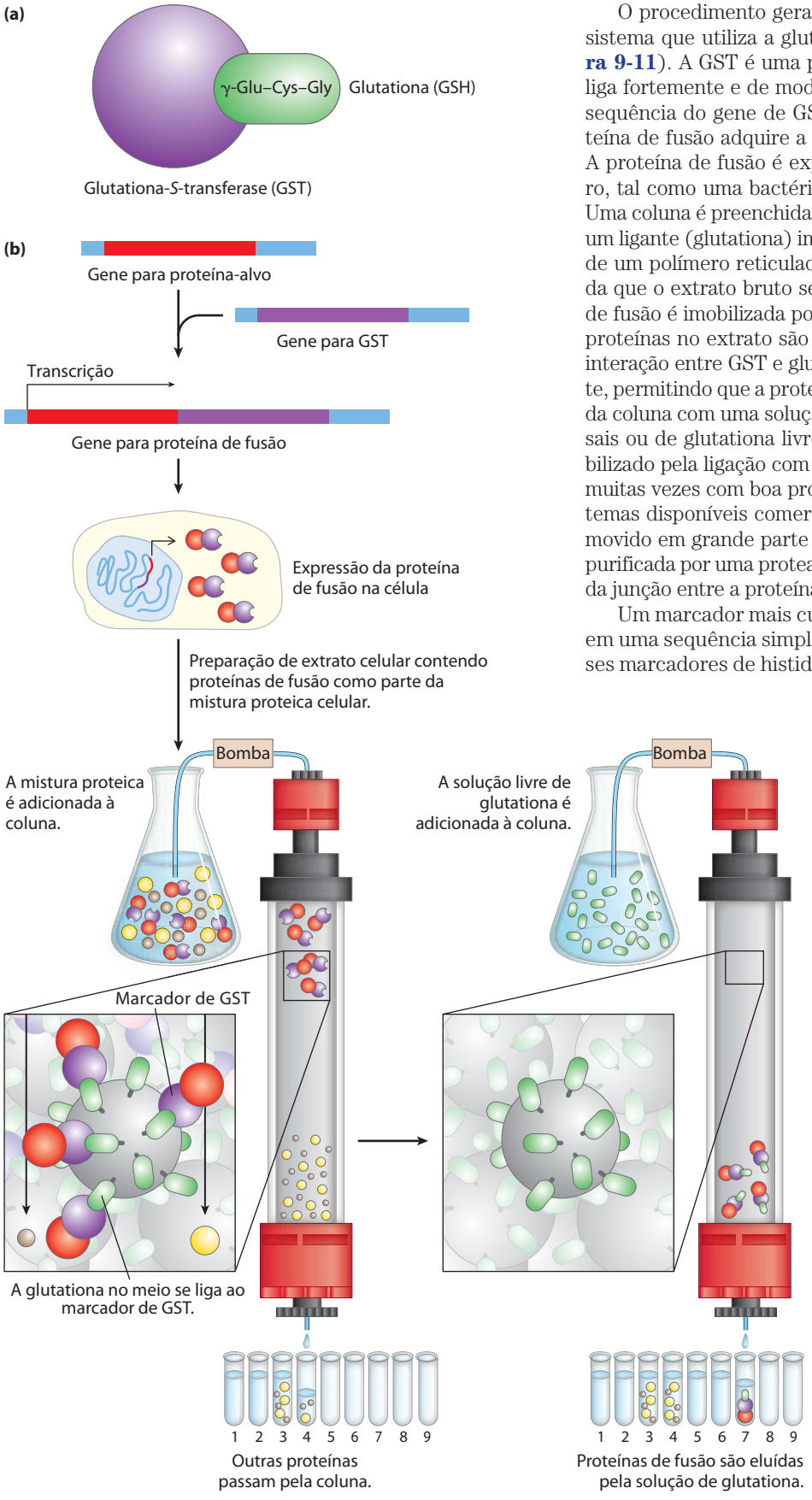


FIGURA 9-11 Utilização de proteínas marcadas na purificação de proteínas. (a) A glutationa-S-transferase (GST) é uma pequena enzima que liga a glutationa (resíduo de glutamato no qual um dipeptídeo Cys-Gly se liga ao carbono da carboxila da cadeia lateral de Glu, daí sua abreviação GSH). **(b)** O marcador de GST se fusiona ao terminal carboxila da proteína por meio de engenharia genética. A proteína marcada expressa na célula está presente no extrato bruto quando as células são lisadas. O extrato é submetido à cromatografia de afinidade (ver Figura 3-17c) por uma matriz com glutationa imobilizada. A proteína marcada com GST se liga à glutationa, retardando sua migração pela coluna, enquanto as outras proteínas são lavadas rapidamente. Depois a proteína marcada é eluída com uma solução contendo uma concentração elevada de sal ou livre de glutationa.

modo firme e especificamente, aos íons de níquel. A matriz da cromatografia com íons de Ni^{2+} imobilizados pode ser utilizada para separar rapidamente uma proteína marcada com His de outras proteínas no extrato. Alguns dos marcadores maiores, como a proteína de ligação de maltose, proporcionam estabilidade e solubilidade adicionais, permitindo a purificação de proteínas clonadas que, de outro modo, seriam inativas em consequência de um enovelamento inadequado ou insolubilidade.

A cromatografia de afinidade utilizando marcadores terminais é um método poderoso e conveniente. Os marcadores têm sido utilizados com sucesso em milhares de estudos publicados; em muitos casos, seria impossível purificar e estudar a proteína sem o marcador. Entretanto, mesmo os marcadores muito pequenos afetam as propriedades das proteínas ligadas a eles, de modo a influenciar os resultados do estudo. Por exemplo, um marcador pode afetar adversamente o enovelamento de proteínas. Mesmo se o marcador for removido por uma protease, um ou alguns resíduos de aminoácidos adicionais ficam por trás da proteína-alvo, afetando ou não a atividade da proteína. Os tipos de experimentos a serem realizados, e os resultados obtidos a partir deles, devem ser sempre avaliados com o auxílio de controles bem desenhados para avaliar o efeito de um marcador no funcionamento de uma proteína.

As sequências gênicas podem ser amplificadas com a reação em cadeia da polimerase

Os projetos do genoma mundial geram bancos de dados internacionais que contêm as sequências genômicas completas de centenas de organismos e proporcionam um acesso sem precedentes à informação da sequência gênica. Essa, por sua vez, simplifica o processo de clonagem de genes individuais para análises mais detalhadas. Se a sequência de, pelo menos, porções da extremidade de um segmento de DNA de interesse é conhecida, o número de cópias do segmento de DNA pode ser muito amplificado com a **reação em cadeia da polimerase** (**PCR**), processo concebido por Kary Mullis em 1983. O DNA amplificado pode então ser clonado por métodos descritos anteriormente ou pode ser utilizado em vários procedimentos analíticos.

O procedimento de PCR tem uma simplicidade elegante. Baseia-se em DNA-polimerases que sintetizam fitas de DNA a partir de desoxirribonucleotídeos, utilizando um molde de DNA (Capítulo 25). Todas as DNA-polimerases sintetizam DNA na direção 5′→ 3′ (ver Figura 8-33a). Além disso, as DNA-polimerases não sintetizam DNA novo, mas, ao contrário, devem adicionar nucleotídeos a fitas preexistentes, denominadas oligonucleotídeos iniciadores (*primers*). Dois oligonucleotídeos sintéticos são preparados, complementares às sequências em fitas opostas ao DNA-alvo, em posições que definem as extremidades do segmento a ser amplificado. Os oligonucleotídeos servem como iniciadores de replicação que podem ser estendidos por uma DNA-polimerase. As extremidades 3′ dos iniciadores hibridizados são orientadas uma para a outra e posicionadas para a síntese do DNA pelo segmento de DNA (**Figura 9-12a**). A PCR básica necessita de quatro componentes: uma amostra de DNA contendo o segmento a ser amplificado, o par de oligonucleotídeos iniciadores sintéticos, desoxinucleotídeos trifosfatos (dNTPs) e DNA-polimerase. A mistura de reação é aquecida brevemente para desnaturar o DNA, separando as duas fitas. A mistura é resfriada de modo a que os oligonucleotídeos iniciadores possam renaturar com o DNA. A concentração elevada de oligonucleotídeos iniciadores aumenta a probabilidade de renaturação com cada fita do DNA desnaturado antes que as duas fitas de DNA (presentes em concentração muito menor) possam renaturar entre si. O segmento é, em seguida, replicado seletivamente pela DNA-polimerase, utilizando o conjunto de dNTPs. O ciclo de aquecimento, resfriamento e replicação é repetido 25 a 30 vezes em algumas horas, em um processo automatizado, amplificando o segmento de DNA entre os oligonucleotídeos iniciadores, até que ele possa ser rapidamente analisado ou clonado. Cada ciclo multiplica a quantidade de segmento de DNA por um fator de 2, de modo que a concentração do DNA cresça exponencialmente. Após 20 ciclos, o segmento de DNA foi amplificado mais de um milhão de vezes (2^{20}); após 30 ciclos, mais de um bilhão de vezes. Todo o DNA restante na amostra permanece sem amplificação. A PCR utiliza uma DNA-polimerase termoestável, que se mantém ativa após cada etapa de aquecimento e não precisa ser reabastecida.

Por meio de um desenho cuidadoso dos oligonucleotídeos iniciadores utilizados para a PCR, o segmento amplificado pode ser modificado pela inclusão, em cada extremidade, de DNA adicional que não está presente no cromossomo-alvo. Por exemplo, sítios de clivagem para endonucleases de restrição podem ser incluídos para facilitar a posterior clonagem do DNA amplificado (Figura 9-12b).

Essa tecnologia é muito sensível: a PCR pode detectar e amplificar tão pouco quanto uma molécula de DNA em quase todo tipo de amostra. A estrutura em dupla-hélice do DNA torna a molécula altamente estável e, de fato, o DNA se degrada lentamente ao longo do tempo (ver Capítulo 8). Entretanto, a PCR permitiu a clonagem bem-sucedida de segmentos de DNA raros e não degradados de amostras de mais de 40.000 anos. Os pesquisadores utilizaram a técnica para clonar fragmentos de DNA dos restos mumificados de seres humanos e animais extintos, como o mamute, criando os novos campos da arqueologia e paleontologia molecular. O DNA de locais de sepultamento foi amplificado por PCR e utilizado para rastrear migrações de humanos antigos. Epidemiologistas podem utilizar amostras de DNA de remanescentes humanos potencializadas por PCR para rastrear a evolução de vírus patogênicos humanos. Assim, além da sua utilidade para a clonagem de DNA, a PCR é uma ferramenta poderosa na medicina forense (Quadro 9-1). Ela pode ser utilizada para detectar infecções virais antes que provoquem sintomas. A PCR é uma ferramenta importante para o aconselhamento genético de potenciais pais de família com condições genéticas hereditárias importantes.

Dada a extrema sensibilidade dos métodos de PCR, a contaminação de amostras se constitui em um sério problema. Em muitas aplicações, incluindo os testes de DNA forense e de arqueologia, as amostras controle asseguram que o DNA amplificado não seja derivado do pesquisador ou de bactérias contaminantes.

Muitas adaptações especializadas de PCR aumentaram a utilidade do método. Por exemplo, as sequências no RNA podem ser amplificadas se o primeiro ciclo de PCR utilizar

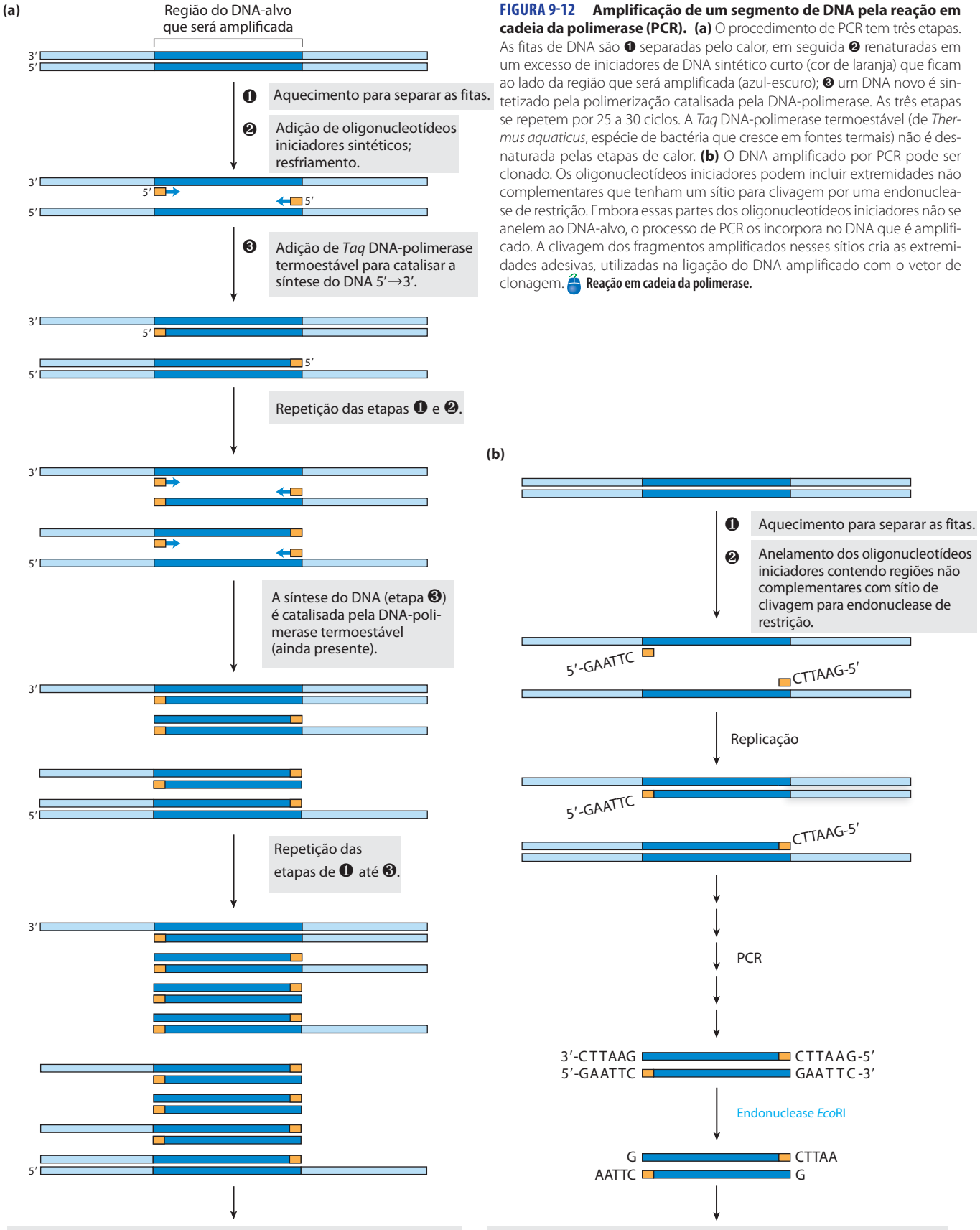


FIGURA 9-12 Amplificação de um segmento de DNA pela reação em cadeia da polimerase (PCR). (a) O procedimento de PCR tem três etapas. As fitas de DNA são ❶ separadas pelo calor, em seguida ❷ renaturadas em um excesso de iniciadores de DNA sintético curto (cor de laranja) que ficam ao lado da região que será amplificada (azul-escuro); ❸ um DNA novo é sintetizado pela polimerização catalisada pela DNA-polimerase. As três etapas se repetem por 25 a 30 ciclos. A *Taq* DNA-polimerase termoestável (de *Thermus aquaticus*, espécie de bactéria que cresce em fontes termais) não é desnaturada pelas etapas de calor. **(b)** O DNA amplificado por PCR pode ser clonado. Os oligonucleotídeos iniciadores podem incluir extremidades não complementares que tenham um sítio para clivagem por uma endonuclease de restrição. Embora essas partes dos oligonucleotídeos iniciadores não se anelem ao DNA-alvo, o processo de PCR os incorpora no DNA que é amplificado. A clivagem dos fragmentos amplificados nesses sítios cria as extremidades adesivas, utilizadas na ligação do DNA amplificado com o vetor de clonagem. **Reação em cadeia da polimerase.**

QUADRO 9-1 MÉTODOS Poderosa ferramenta na medicina legal

Um dos métodos mais precisos para a colocação de um indivíduo na cena de um crime é a impressão digital. O advento da tecnologia do DNA recombinante disponibilizou uma ferramenta muito mais poderosa: a **genotipagem de DNA** (também chamada de DNA *fingerprinting* ou perfis de DNA). O método se baseia nos polimorfismos da sequência – diferenças sutis na sequência entre os indivíduos: um em cada 1.000 pb, em média. O uso dessas diferenças para identificação forense foi descrito pela primeira vez pelo geneticista inglês Alec Jeffreys em 1985. Cada diferença em relação ao protótipo de sequência do genoma humano (o primeiro obtido) ocorre em alguma fração na população humana; cada pessoa tem algumas diferenças em relação a esse protótipo.

O trabalho forense moderno concentra-se em diferenças nos comprimentos das sequências **repetições curtas em tandem** (**STR**). A STR é uma sequência de DNA curta, repetida muitas vezes em tandem em um local específico em um cromossomo; geralmente, a sequência repetida tem 4 pb de comprimento. Os *loci* STR mais frequentemente utilizados na genotipagem de DNA são curtos – 4 a 50 repetições de comprimento (16 a 200 pb para repetições de tetranucleotídeos) – e têm múltiplas variantes de comprimento na população humana. Mais de 20.000 *loci* STR de tetranucleotídeos foram caracterizados e mais de um milhão de STR de todos os tipos podem estar presentes no genoma humano, o que representa cerca de 3% de todo o DNA humano.

(Continua na próxima página)

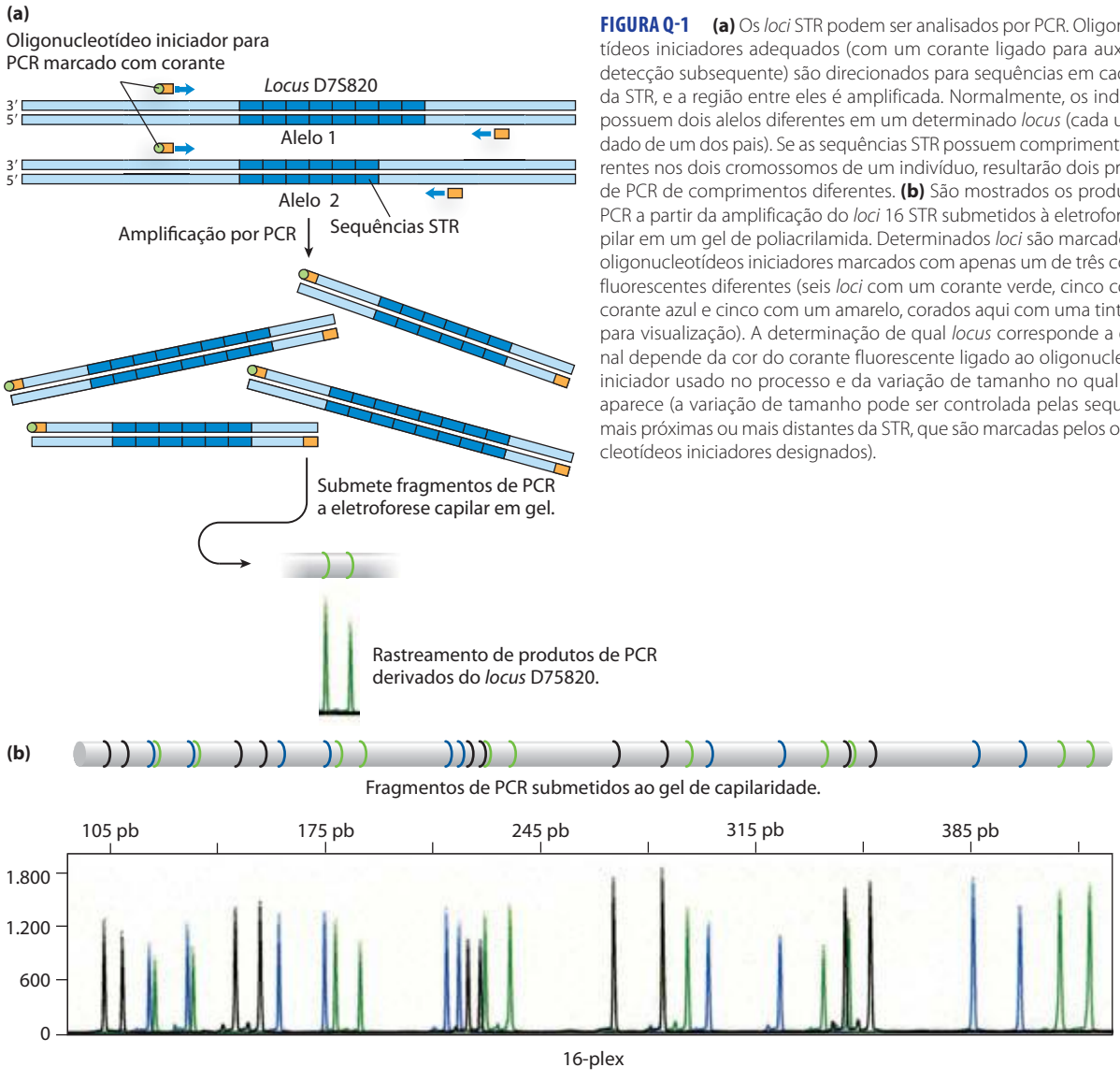


FIGURA Q-1 **(a)** Os *loci* STR podem ser analisados por PCR. Oligonucleotídeos iniciadores adequados (com um corante ligado para auxiliar na detecção subsequente) são direcionados para sequências em cada lado da STR, e a região entre eles é amplificada. Normalmente, os indivíduos possuem dois alelos diferentes em um determinado *locus* (cada um herdado de um dos pais). Se as sequências STR possuem comprimentos diferentes nos dois cromossomos de um indivíduo, resultarão dois produtos de PCR de comprimentos diferentes. **(b)** São mostrados os produtos de PCR a partir da amplificação do *loci* 16 STR submetidos à eletroforese capilar em um gel de poliacrilamida. Determinados *loci* são marcados com oligonucleotídeos iniciadores marcados com apenas um de três corantes fluorescentes diferentes (seis *loci* com um corante verde, cinco com um corante azul e cinco com um amarelo, corados aqui com uma tinta preta para visualização). A determinação de qual *locus* corresponde a qual sinal depende da cor do corante fluorescente ligado ao oligonucleotídeo iniciador usado no processo e da variação de tamanho no qual o sinal aparece (a variação de tamanho pode ser controlada pelas sequências, mais próximas ou mais distantes da STR, que são marcadas pelos oligonucleotídeos iniciadores designados).

QUADRO 9-1 MÉTODOS Poderosa ferramenta na medicina legal *(Continuação)*

O comprimento de uma STR particular em um indivíduo específico pode ser determinado com o auxílio da reação em cadeia da polimerase (ver Figura 9-12). O uso de PCR também torna o procedimento sensível o suficiente para ser aplicado a amostras muito pequenas de DNA recolhidas muitas vezes em cenas de crime. As sequências de DNA que acompanham as STR são exclusivas para cada *locus* STR e idênticas (exceto para mutações extremamente raras) em todos os humanos. Os oligonucleotídeos iniciadores da PCR são direcionados para esse DNA que acompanha a STR e são projetados para amplificar o DNA ao longo da STR (Figura Q-1a). O comprimento do produto da PCR, portanto, reflete o comprimento da STR nessa amostra. Como cada ser humano herda um cromossomo de cada par de cromossomos de cada progenitor, os comprimentos de STR nos dois cromossomos são frequentemente diferentes, gerando dois comprimentos de STR diferentes a partir de um indivíduo. Os produtos de PCR são submetidos à eletroforese em gel de poliacrilamida muito fino em um tubo capilar. As bandas resultantes são convertidas em um conjunto de picos que revelam com precisão o tamanho de cada fragmento de PCR e, assim, o comprimento da STR no alelo correspondente. A análise de múltiplos *loci* STR pode produzir um perfil exclusivo para um indivíduo (Figura Q-1b). Isso normalmente é realizado com um *kit* disponível comercialmente que inclui oligonucleotídeos iniciadores de PCR únicos para cada *locus*, ligados a corantes coloridos para ajudar a distinguir os diferentes produtos de PCR. A amplificação por PCR permite aos investigadores obter genótipos de STR de menos de 1 ng de DNA parcialmente degradado, quantidade que pode ser obtida a partir de um único folículo piloso, uma gota de sangue, uma pequena amostra de sêmen ou amostras com meses ou até mesmo vários anos de idade. Quando bons genótipos de STR são obtidos, a possibilidade de erros de identificação é inferior a 1 em 10^{18} (quintilhão).

A utilização forense bem-sucedida da análise de STR necessitou de padronização, conseguida pela primeira vez no Reino Unido em 1995. O padrão dos EUA, Combined DNA Index System (CODIS, sistema combinado do índice do DNA), criado em 1998, baseia-se em 13 *loci* STR bem estudados, que devem estar presentes em qualquer experimento de tipagem de DNA realizado nos Estados Unidos (Tabela Q-1). O gene amelogenina também é utilizado como marcador nas análises. Presente em cromossomos sexuais humanos, esse gene tem um comprimento um pouco diferente nos cromossomos X e Y. Assim, a amplificação por PCR por meio desse gene gera produtos de diferentes tamanhos que revelam o sexo do doador de DNA. Até o início de 2010, o banco de dados CODIS continha mais de 7 milhões de genótipos de STR e tinha auxiliado mais de 100.000 investigações forenses.

A genotipagem de DNA é utilizada tanto para condenar ou para absolver suspeitos e para estabelecer a paternidade com um extraordinário grau de certeza. O impacto desses procedimentos em casos de tribunal continuará a crescer à medida que os padrões forem aprimorados e os bancos de dados internacionais de genotipagem STR aumentarem. Mesmo casos misteriosos muito antigos poderão ser resolvidos. Em 1996, o *fingerprinting* de DNA ajudou a confirmar a identificação dos ossos do último czar russo e de sua família, que foram assassinados em 1918.

TABELA Q-1 Propriedades dos *loci* utilizados para o banco de dados CODIS

Locus	Cromossomo	Motivo de repetição	Comprimento da repetição (média)*	Número de alelos observados†
CSF1PO	5	TAGA	5–16	20
FGA	4	CTTT	12,2–51,2	80
TH01	11	TCAT	3–14	20
TPOX	2	GAAT	4–16	15
VWA	12	[TCTG][TCTA]	10–25	28
D3S1358	3	[TCTG][TCTA]	8–21	24
D5S818	5	AGAT	7–18	15
D7S820	7	GATA	5–16	30
D8S1179	8	[TCTA][TCTG]	7–20	17
D13S317	13	TATC	5–16	17
D16S539	16	GATA	5–16	19
D18S51	18	AGAA	7–39,2	51
D21S11	21	[TCTA][TCTG]	12–41,2	82
Amelogenin‡	X, Y	Não aplicado		

Fonte: Adaptado a partir de Butter, J.M. (2005) *Forensic DNA Typing*, 2nd edn, Academic Press, San Diego, p. 96.
*Comprimentos repetidos observados na população humana. As repetições parciais ou imperfeitas podem ser incluídas em alguns alelos.
†Número de diferentes alelos observados desde 2005 na população humana. A análise cuidadosa de um *locus* em vários indivíduos é um pré-requisito para seu uso na tipagem forense do DNA.
‡A amelogenina é um gene, de tamanho levemente diferente no cromossomo X e Y, utilizado para estabelecer o gênero.

a transcriptase reversa, uma enzima que funciona como uma DNA-polimerase (ver Figura 8-33a), mas utiliza o RNA como molde (Figura 9-12). Após a síntese da fita de DNA a partir do molde de RNA, os ciclos restantes podem ser feitos com DNA-polimerases, utilizando protocolos padrão de PCR. Essa **PCR pela transcriptase reversa** (**RT-PCR**) pode ser utilizada, por exemplo, para detectar sequências derivadas de células vivas (que fazem a transcrição de seu DNA em RNA) em oposição a tecidos mortos.

Protocolos de PCR também podem ser quantitativos para estimar o número de cópias relativas a sequências específicas em uma amostra. O método é chamado de **PCR quantitativa**, ou **qPCR**. Se uma sequência de DNA estiver presente em quantidades maiores que as habituais em uma amostra – por exemplo, se alguns genes são amplificados em células tumorais –, a qPCR pode revelar o aumento da representação daquela sequência. Em resumo, a PCR é realizada na presença de uma sonda que emite um sinal fluorescente quando o produto de PCR está presente (**Figura 9-13**). Se a sequência de interesse está presente em níveis mais elevados do que em outras sequências na amostra, o sinal de PCR irá atingir um limiar predeterminado mais rápido. A PCR de transcriptase reversa e a PCR em tempo real podem ser combinadas para determinar as concentrações relativas de uma molécula específica de RNA em uma célula em diferentes condições ambientais.

Finalmente, a PCR facilitou o desenvolvimento de novos procedimentos poderosos de sequenciamento de DNA, como será visto.

RESUMO 9.1 Estudo dos genes e seus produtos

- A clonagem do DNA e a engenharia genética envolvem a clivagem do DNA e o conjunto de segmentos de DNA em novas combinações – DNA recombinante.
- A clonagem envolve cortar o DNA em fragmentos com enzimas; selecionar e possivelmente modificar um fragmento de interesse; inserir o fragmento de DNA em um vetor de clonagem adequado; transferir o vetor com o DNA inserido em uma célula hospedeira para a replicação; e identificar e selecionar células que contêm o fragmento de DNA.
- Enzimas essenciais na clonagem gênica incluem endonucleases de restrição (especialmente as enzimas tipo II) e a DNA-ligase.
- Vetores de clonagem incluem plasmídeos e, para as inserções de DNA mais longas, cromossomos artificiais de bactérias (BAC) e cromossomos artificiais de leveduras (YAC).
- Técnicas de engenharia genética manipulam células para expressar e/ou alterar genes clonados.
- As proteínas ou peptídeos podem se ligar a uma proteína de interesse alterando o seu gene clonado, criando uma proteína de fusão. Os segmentos de peptídeos adicionais podem ser utilizados para detectar a proteína ou purificá-la utilizando métodos de cromatografia de afinidade convenientes.
- A reação em cadeia da polimerase (PCR) permite a amplificação de segmentos escolhidos de DNA ou RNA para estudo detalhado ou clonagem.

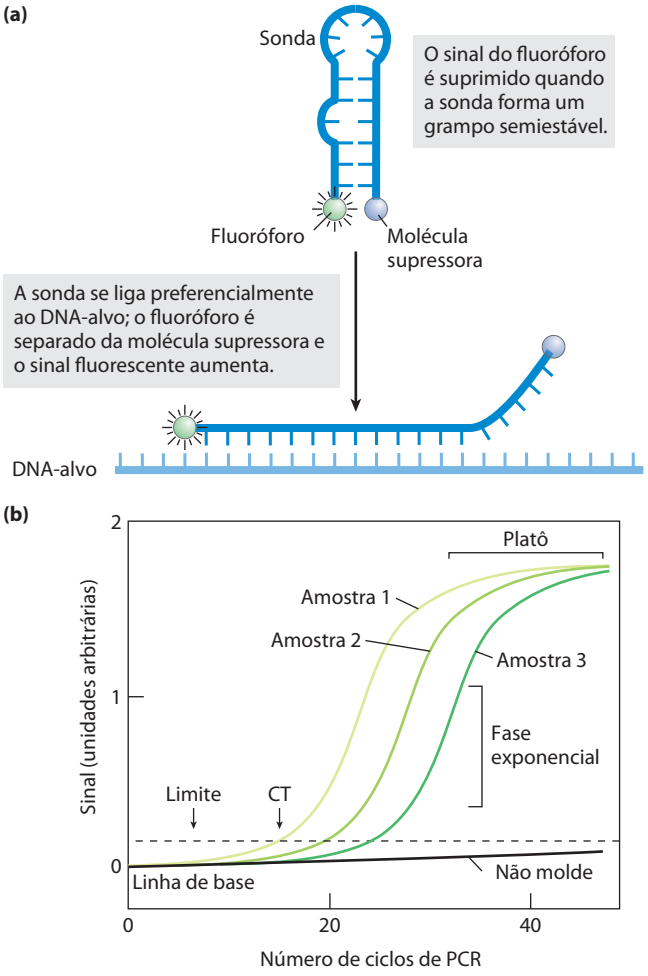


FIGURA 9-13 PCR quantitativa. A PCR pode ser utilizada quantitativamente, monitorando-se cuidadosamente o progresso da amplificação da PCR e determinando quando um segmento de DNA foi amplificado para um nível limite específico. **(a)** A quantidade de produto de PCR presente é determinada medindo-se o nível de uma sonda fluorescente ligada a um oligonucleotídeo repórter complementar ao segmento de DNA que está sendo amplificado. A fluorescência da sonda não é detectada inicialmente porque um supressor de fluorescência é ligado ao mesmo nucleotídeo. Quando o oligonucleotídeo repórter se pareia com seu complemento em uma cópia do segmento de DNA amplificado, o fluoróforo se separa da molécula supressora e a fluorescência aparece. **(b)** À medida que prossegue a reação de PCR, a quantidade de segmento de DNA-alvo aumenta exponencialmente, e o sinal fluorescente também aumenta exponencialmente à medida que as sondas de oligonucleotídeos se ligam aos segmentos amplificados. Após vários ciclos de PCR, o sinal alcança um platô à medida que um ou mais componentes da reação se extinguem. Quando um segmento está presente em quantidades maiores em uma amostra do que em outra, sua amplificação atinge um limiar predefinido. A linha "Não molde" segue o lento aumento no sinal de fundo observado em um controle que não inclui uma amostra de DNA adicionada. CT é o número de ciclos no qual o limiar é ultrapassado primeiro.

9.2 Utilização de métodos com base no DNA para a compreensão das funções das proteínas

A função das proteínas pode ser descrita em três níveis. A **função fenotípica** descreve os efeitos de uma proteína em todo o organismo. Por exemplo, a perda de proteínas pode levar a um crescimento mais lento do organismo, a uma alteração no padrão de desenvolvimento ou até mesmo

à morte do organismo. A **função celular** é a descrição de uma rede de interações em que as proteínas participam no nível celular. A identificação de interações com outras proteínas na célula pode ajudar a definir os tipos de processos metabólicos nos quais a proteína participa. Finalmente, a **função molecular** refere-se à atividade bioquímica precisa de uma proteína, incluindo detalhes como as reações que uma enzima catalisa ou os ligantes que ligam a um receptor. O desafio de compreender as funções de milhares de proteínas não caracterizadas ou mal caracterizadas encontradas em uma célula comum originou uma grande variedade de técnicas. Os métodos com base no DNA são uma contribuição fundamental para esse esforço e fornecem informações em todos os três níveis.

Bibliotecas de DNA são catálogos especializados de informação genética

Uma **biblioteca de DNA** é uma coleção de clones de DNA reunidos para fins de sequenciamento do genoma, descoberta de genes ou a determinação da função de um gene/proteína. A biblioteca pode ter diversas formas, dependendo da fonte de DNA e do propósito final da biblioteca.

A maior é uma **biblioteca genômica**, produzida quando o genoma completo de um organismo é clivado em milhares de fragmentos. Todos os fragmentos são clonados por inserção de cada fragmento em um vetor de clonagem. Isso cria uma mistura complexa de vetores recombinantes, cada um com um fragmento clonado diferente. A primeira etapa é a digestão *parcial* do DNA por endonucleases de restrição, de forma que qualquer sequência apareça em fragmentos em um número limitado de tamanhos – uma faixa de tamanhos compatível com o vetor de clonagem para assegurar que praticamente todas as sequências estejam representadas entre os clones na biblioteca. Os fragmentos muito grandes ou muito pequenos para a clonagem são removidos por centrifugação ou eletroforese. O vetor de clonagem, como BAC ou YAC, é clivado com a mesma endonuclease de restrição utilizada para digerir o DNA e ligado aos fragmentos de DNA genômico. A mistura de DNA ligado é então utilizada para transformar células bacterianas ou de leveduras para produzir uma biblioteca de células, cada qual abrigando uma molécula diferente de DNA recombinante. Idealmente, todo o DNA no genoma em estudo é representado na biblioteca. Cada célula bacteriana ou de levedura cresce em uma colônia ou clone de células idênticas, cada célula com o mesmo plasmídeo recombinante, um dos muitos representados na biblioteca total.

Os esforços para definir a função de um gene ou proteína muitas vezes se utilizam de bibliotecas mais especializadas. Um exemplo é uma biblioteca que inclui apenas as sequências de DNA *expressas*, isto é, transcritas em um RNA – em determinado organismo, ou mesmo em apenas algumas células ou tecidos. Essa biblioteca não tem o DNA não codificante que faz parte de grande parte de muitos genomas eucarióticos. Em primeiro lugar, o pesquisador extrai o mRNA de um organismo ou de células específicas de um organismo e, em seguida, prepara os **DNA complementares** (**cDNA**). Essa reação em múltiplas etapas, mostrada na **Figura 9-14**, depende da transcriptase reversa, que sintetiza DNA a partir de um molde de RNA. Os fragmentos de DNA de cadeia dupla resultante são inseridos num vetor adequado e clonados, criando uma população de clones chamada de **biblioteca de cDNA**. O aparecimento de um gene para uma determinada proteína nessa biblioteca implica que ele é expresso nas células e sob as condições utilizadas para gerar a biblioteca.

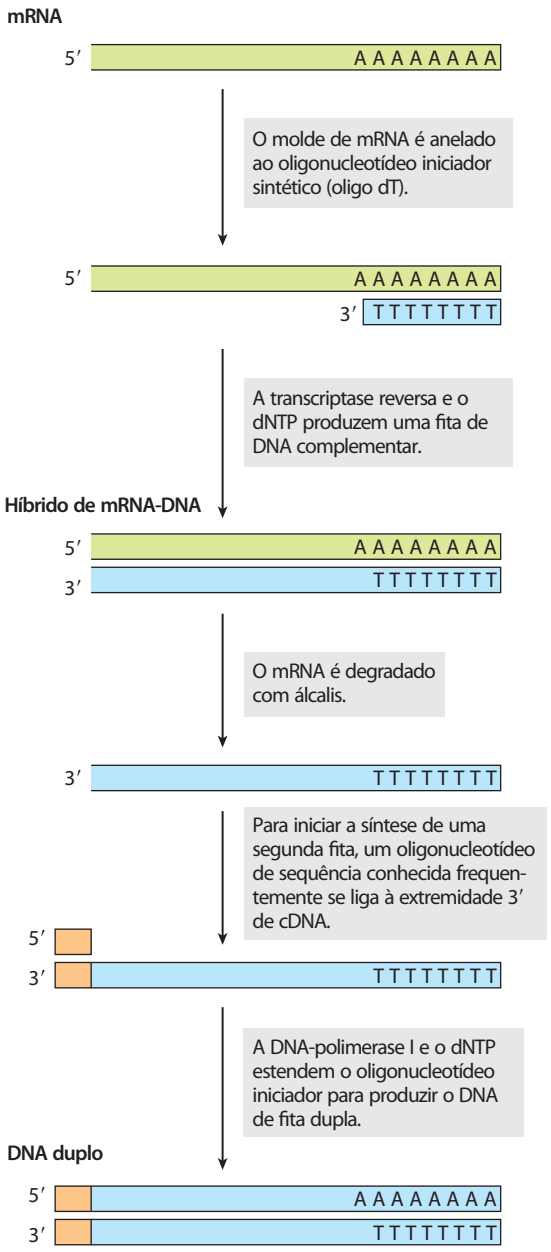


FIGURA 9-14 Construindo uma biblioteca de cDNA a partir de mRNA. O conteúdo total de mRNA de uma célula inclui os transcritos de milhares de genes, e os cDNAs gerados desse mRNA são correspondentemente heterogêneos. A transcriptase reversa pode sintetizar DNA em um molde de RNA ou DNA (ver Figura 26-32). Para iniciar a síntese da segunda fita de DNA, os oligonucleotídeos de sequência conhecida se ligam à extremidade 3′ da primeira fita, e o cDNA de fita dupla então produzido é clonado em um plasmídeo.

Sequências ou relações estruturais fornecem informações sobre a função das proteínas

Uma razão importante para se sequenciar muitos genomas é fornecer um banco de dados capaz de atribuir funções de genes por comparações do genoma, empreendimento conhecido como **genômica comparativa**. Algumas vezes, um gene recém-descoberto é relatado por homologias de sequências de um gene previamente estudado em outra ou na mesma espécie, e a sua função pode ser total ou parcialmente definida por essa relação. Os genes que ocorrem em espécies diferentes, mas têm uma sequência e uma relação funcional clara entre si, são chamados **ortólogos**. Genes com relações semelhantes entre si, dentro de uma única espécie, são chamados de **parálogos**. Se a função de um gene foi caracterizada para uma espécie, essa informação pode ser utilizada para, ao menos, atribuir provisoriamente a função do gene para o ortólogo encontrado em uma segunda espécie. A correlação é mais fácil de ser realizada quando se comparam genomas de espécies com relativa proximidade, como camundongos e humanos, embora muitos genes claramente ortólogos tenham sido identificados em espécies tão distantes quanto as bactérias e os seres humanos. Algumas vezes, até mesmo a ordem dos genes em um cromossomo é conservada ao longo de grandes segmentos de genomas de espécies estreitamente relacionadas (**Figura 9-15**). A ordem gênica conservada, chamada **sintenia**, fornece evidência adicional para uma relação ortóloga entre genes em locais idênticos dentro dos segmentos relacionados.

Alternativamente, algumas sequências associadas a determinados motivos estruturais (Capítulo 4) podem ser identificadas em uma proteína. A presença de um motivo estrutural pode ajudar a definir a função molecular, sugerindo que uma proteína, por exemplo, catalisa a hidrólise de ATP, liga-se ao DNA ou forma um complexo com íons de zinco. Essas relações são determinadas com a ajuda de programas de computador sofisticados, limitados apenas pela informação atual sobre a estrutura do gene e da proteína e pela capacidade para associar sequências com motivos estruturais determinados.

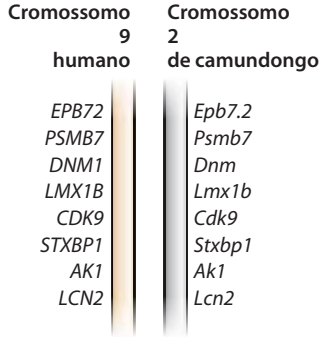


FIGURA 9-15 Sintenia em genomas humanos e de camundongos. Segmentos grandes dos dois genomas têm genes intimamente relacionados alinhados na mesma ordem nos cromossomos. Nos segmentos curtos do cromossomo 9 humano e no cromossomo 2 de camundongo, os genes exibem alto grau de homologia, assim como a mesma ordem no gene. Os diferentes esquemas de letras para os nomes dos genes simplesmente refletem as diferentes convenções de nomenclatura nas duas espécies.

Proteínas de fusão e imunofluorescência podem localizar proteínas em células

Muitas vezes, uma pista importante para a função de um produto gênico surge pela determinação de sua localização no interior da célula. Por exemplo, uma proteína que se encontra exclusivamente no núcleo pode estar envolvida em processos exclusivos dessa organela, tais como transcrição, replicação ou condensação da cromatina. Com frequência, pesquisadores modificam geneticamente proteínas de fusão com a finalidade de localizar uma proteína na célula ou organismo. Algumas das fusões mais úteis envolvem a ligação de proteínas marcadoras que permitem ao investigador determinar a localização por visualização direta ou por imunofluorescência.

Um marcador particularmente útil é a **proteína fluorescente verde** (**GFP**). Um gene-alvo (que codifica a proteína de interesse) fusionado ao gene de GFP gera uma proteína de fusão altamente fluorescente – ela literalmente acende quando exposta à luz azul – e pode ser visualizada diretamente em uma célula viva. A GFP é uma proteína derivada da água-viva *Aequorea victoria* (ver Quadro 12-3, Figura 1). Ela tem estrutura em barril β e o fluoróforo (o componente fluorescente da proteína) está no centro do barril (**Figura 9-16a**). O fluoróforo é derivado de um rearranjo e oxidação de vários resíduos de aminoácidos. Como essa reação é autocatalítica e não exige outras proteínas ou outros cofatores além do oxigênio molecular (ver Quadro 12-3, Figura 3), a GFP é prontamente clonada em uma forma ativa em quase todas as células. Apenas algumas moléculas dessa proteína podem ser observadas ao microscópio, permitindo o estudo de sua localização e movimentos em uma célula. Uma cuidadosa modificação genética da proteína, associada ao isolamento de proteínas fluorescentes relacionadas provenientes de outros celenterados marinhos, disponibilizou uma ampla variedade dessas proteínas, em um conjunto de cores (Figura 9-16b) e outras características (brilho, estabilidade). Com essa tecnologia, por exemplo, a proteína GLR1 (receptor de glutamato do tecido nervoso) foi visualizada como proteína de fusão GLR1-GFP no nematódeo *Caenorhabditis elegans* (Figura 9-16c).

Em muitos casos, a visualização de uma proteína de fusão GFP em uma célula viva real não é possível, prática ou desejável. A proteína de fusão GFP pode estar inativa ou pode não ser expressa em níveis suficientes para permitir a visualização. Nesse caso, a **imunofluorescência** é um método alternativo para a visualização da proteína endógena (inalterada). Esse método requer a fixação (e, portanto, a morte) da célula. A proteína de interesse é algumas vezes expressa como uma proteína de fusão com um **marcador de epítopo**, sequência proteica curta que se liga firmemente a um anticorpo bem caracterizado e disponível comercialmente. As moléculas fluorescentes (fluorocromos) estão ligadas a esse anticorpo. Mais comumente, a proteína-alvo permanece inalterada. Essa proteína está ligada por um anticorpo que é específico para ela. Em seguida, um segundo anticorpo é adicionado e se liga especificamente ao primeiro, e é o segundo anticorpo, que tem o(s) fluorocromo(s) ligado(s) (**Figura 9-17a**). Uma variação desse método de visualização indireta é ligar moléculas de biotina ao primeiro anticorpo e, em seguida, adicionar estreptavidina (proteína bacteriana intimamente relaciona-

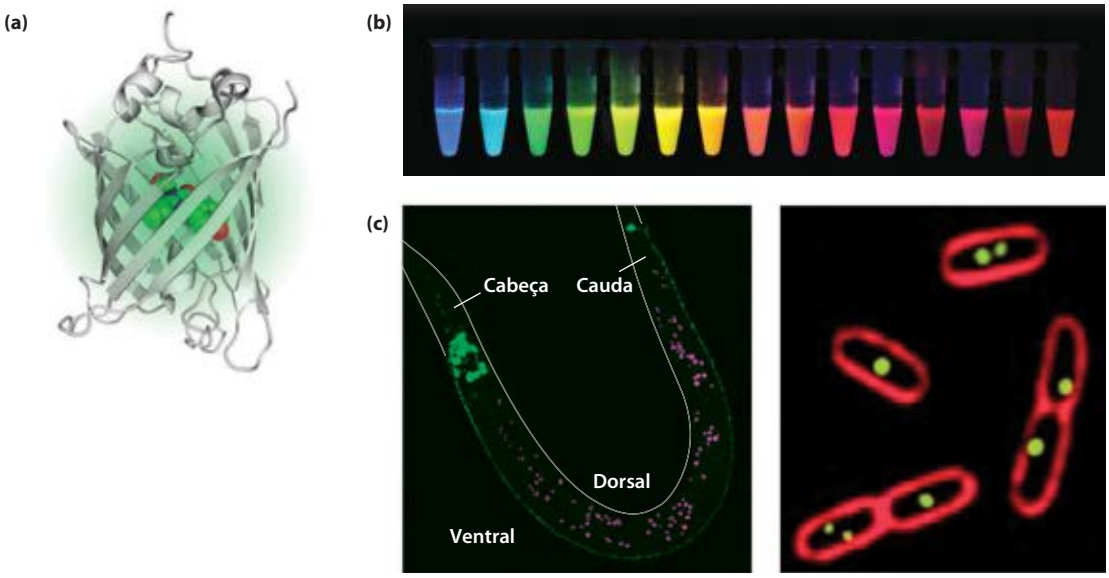


FIGURA 9-16 **Proteína fluorescente verde (GFP).** **(a)** A proteína GFP (PDB ID 1GFL), derivada da água-viva *Aequorea victoria*, tem estrutura em barril β; o fluoróforo (apresentado como modelo de volume atômico) é o centro do barril. **(b)** Variantes de GFP estão disponíveis atualmente em quase todas as cores do espectro visível. **(c)** Uma proteína de fusão GLR1-GFP tem brilho fluorescente verde brilhante no verme nematódeo *Caenorhabditis elegans* (à esquerda). A GLR1 é um receptor de glutamato do tecido nervoso (gotas de gordura autofluorescente se colorem falsamente de magenta). As membranas das células de *E. coli* (à direita) são coradas com corante fluorescente vermelho. As células estão expressando uma proteína que se liga a um plasmídeo residente, fusionado com GFP. Os pontos verdes indicam as posições dos plasmídeos.

da à avidina, proteína que se liga à biotina; ver Tabela 5-1) complexada com fluorocromos. A interação entre a biotina e a estreptavidina é uma das interações mais fortes e mais específicas conhecidas, e o potencial para adicionar múltiplos fluorocromos para cada proteína-alvo confere a esse método uma grande sensibilidade.

Bibliotecas de cDNA altamente especializadas (Figura 9-14) podem ser feitas por clonagem de cDNA ou de fragmentos de cDNA em um vetor que se fusiona a cada sequência de cDNA com a sequência para um marcador chamado de gene repórter. O gene fusionado é frequentemente chamado de construtor repórter. Por exemplo, todos os genes na biblioteca podem ser fusionados com o gene da GFP (**Figura 9-18**). Cada célula na biblioteca expressa um desses genes fusionados. A localização celular do produto de qualquer gene representado na biblioteca será revelada em forma de focos de luz em células que expressam o gene fusionado adequado em níveis suficientes – supondo-se que o gene mantenha a sua função e localização normal.

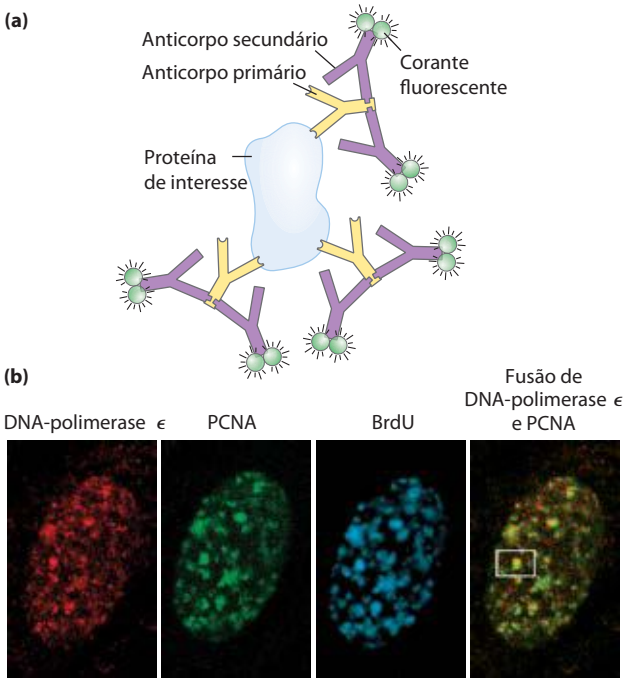


FIGURA 9-17 **Imunofluorescência indireta.** **(a)** A proteína de interesse se liga ao anticorpo primário, e um anticorpo secundário é adicionado; este segundo anticorpo, com um ou mais grupos fluorescentes ligados, se liga ao primeiro. Múltiplos anticorpos secundários podem se ligar ao anticorpo primário, amplificando o sinal. Se a proteína de interesse está no interior de uma célula, a célula é fixada e permeabilizada, e os dois anticorpos são adicionados em sucessão. **(b)** O resultado final é uma imagem na qual os pontos brilhantes indicam a localização da proteína ou proteínas de interesse na célula. As imagens mostram um núcleo de um fibroblasto humano, sucessivamente corado com anticorpos e marcadores fluorescentes para DNA-polimerase ε, para PCNA, importante proteína acessória da polimerase, e para bromodesoxiuridina (BrdU), análogo de nucleotídeo. A BrdU, adicionada como um breve pulso, identifica as regiões de replicação ativa de DNA. Os padrões de coloração mostram que a DNA-polimerase ε e o PCNA se colocalizam em regiões de síntese de DNA ativa. Uma dessas regiões é visível no quadro branco.

Interações proteína-proteína ajudam a elucidar a função de proteínas

Outra chave para definir a função de uma proteína específica é determinar o quê se liga a essa proteína. No caso de interações proteína-proteína, a associação de uma proteína

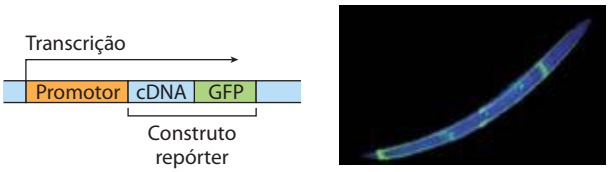


FIGURA 9-18 Bibliotecas especializadas de DNA. A clonagem de um cDNA próximo ao gene de GFP cria um construto repórter. A transcrição prossegue por meio do gene de interesse (o cDNA inserido) e o gene repórter (aqui, o de GFP), e o transcrito de mRNA é expresso como uma proteína de fusão. Parte da proteína GFP é visível ao microscópio com fluorescência. Embora seja mostrado apenas um exemplo, milhares de genes podem ser fusionados com GFP em construtos similares e armazenados em bibliotecas nas quais cada célula ou organismo, na biblioteca, expressa uma proteína diferente fusionada à GFP. Se a proteína de fusão é adequadamente expressa, sua localização na célula ou organismo pode ser avaliada. A fotografia mostra um nematódeo contendo uma proteína de fusão GFP expressa apenas nos quatro neurônios "de toque" que percorrem o comprimento do seu corpo.

de função desconhecida àquela cuja função é conhecida pode fornecer uma "culpa por associação" útil e atraente. As técnicas utilizadas nesse esforço são bastante variadas.

Purificação de complexos de proteínas Com a construção de bibliotecas de cDNA em que cada gene é fusionado a um marcador de epítopo, os investigadores podem precipitar o produto proteico de um gene por meio de sua complexação com o anticorpo que se liga ao epítopo, processo denominado **imunoprecipitação** (**Figura 9-19**). Se a proteína marcada é expressa em células, outras proteínas que se ligam a ela também precipitam com ela. A identificação de proteínas associadas revela algumas das interações proteína-proteína da proteína marcada. Há muitas variações desse processo. Por exemplo, um extrato bruto de células que expressa uma proteína marcada é adicionado a uma coluna contendo anticorpos imobilizados (ver na Figura 3-17c uma descrição de cromatografia de afinidade). A proteína marcada se liga ao anticorpo, e, às vezes, proteínas que interagem com a proteína marcada também são retidas na coluna. A conexão entre a proteína e o marcador é clivada com uma protease específica e os complexos de proteínas são eluídos a partir da coluna e analisados. Os pesquisadores utilizam esses métodos para definir redes complexas de interações no interior de uma célula. Em princípio, o método cromatográfico para analisar interações proteína-proteína é utilizado com qualquer tipo de marcador de proteína (marcador His, GST, etc.) que possa ser imobilizado em meio cromatográfico adequado.

A seletividade desse método foi potencializada com os **marcadores de purificação por afinidade em tandem** (**TAP**). Dois marcadores consecutivos são fusionados a uma proteína-alvo, e a proteína de fusão é expressa em uma célula (**Figura 9-20**). O primeiro marcador é a proteína A, encontrada na superfície da bactéria *Staphylococcus aureus* que se liga firmemente à imunoglobulina G (IgG) de mamíferos. O segundo marcador é frequentemente um peptídeo de ligação de calmodulina. Um extrato bruto contendo a proteína de fusão marcadora de TAP é passado por uma matriz de coluna com anticorpos IgG fixados que se ligam à proteína A. A maior parte das proteínas celulares não ligadas é lavada pela coluna, mas as proteínas que normalmente interagem com a proteína-alvo na célula são retidas. O primeiro marcador é, então, clivado a partir da

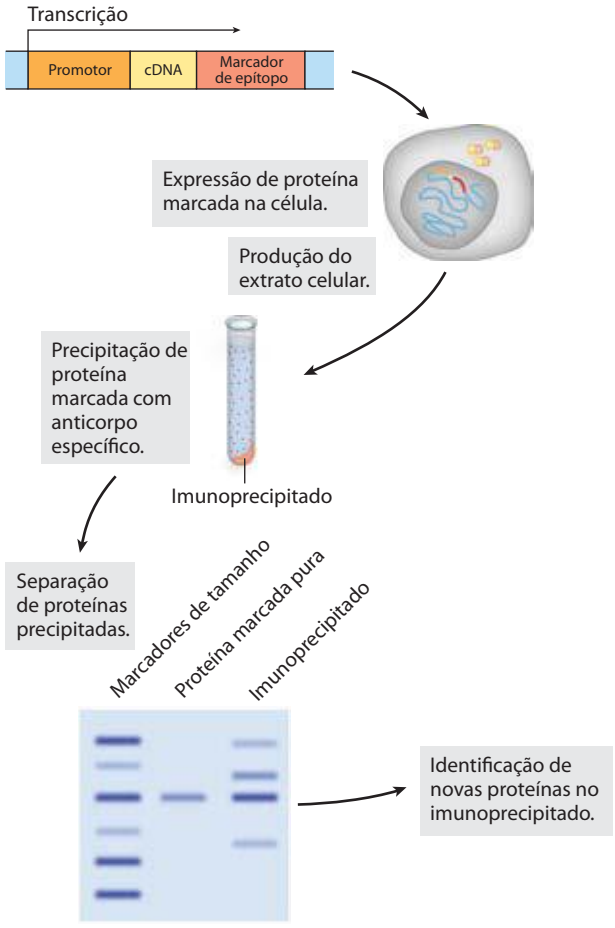


FIGURA 9-19 O uso de marcadores de epítopos para estudar interações proteína-proteína. O gene de interesse é clonado próximo a um gene para um marcador de epítopo, e a proteína de fusão resultante é precipitada por anticorpos para o epítopo. Todas as outras proteínas que interagem com a proteína marcada também precipitam, contribuindo assim para elucidar as interações proteína-proteína.

proteína de fusão com uma protease altamente específica, a protease TEV, e a proteína-alvo de fusão encurtada e quaisquer outras proteínas associadas de modo não covalente à proteína-alvo são eluídas na coluna. O eluente, então, passa por uma segunda coluna contendo uma matriz com calmodulina fixada, que liga o segundo marcador. Proteínas ligadas frouxamente voltam a ser lavadas ao longo da coluna. Após a clivagem do segundo marcador, a proteína-alvo é eluída ao longo da coluna com suas proteínas associadas. As duas etapas de purificação consecutivas eliminam todos os contaminantes fracamente ligados. Falsos-positivos são minimizados e as interações proteicas que persistem em ambas as etapas podem ser funcionalmente significativas.

Análise de duplos-híbridos em leveduras Um método genético sofisticado para definir interações proteína-proteína se baseia nas propriedades da proteína Gal4 (Gal4p; ver Figura 28-30), que ativa a transcrição de genes *GAL* em genes de leveduras, que codificam as enzimas do metabolismo da galactose. A Gal4p tem dois domínios: um que se liga a uma sequência específica de DNA e outro que ativa a RNA-polimerase para sintetizar mRNA a partir de um gene adjacente.

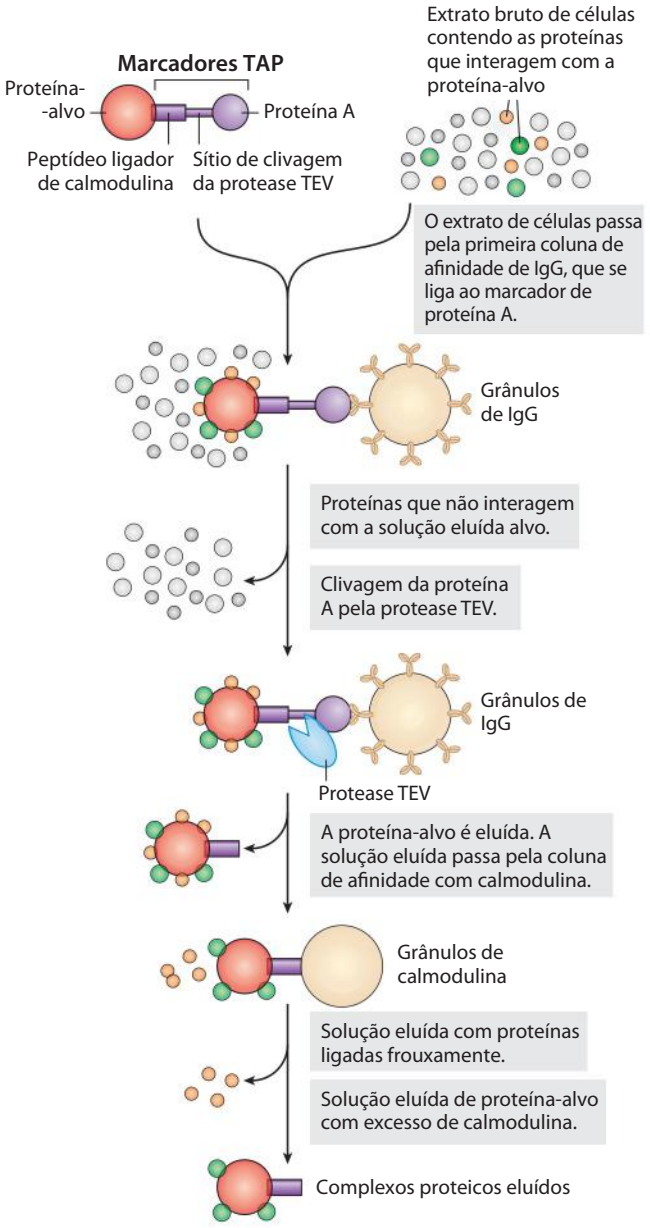


FIGURA 9-20 Marcadores por purificação de afinidade em tandem (TAP). Uma proteína marcada com TAP e proteínas associadas são isoladas por duas purificações de afinidade consecutivas, conforme descrito no texto.

Os dois domínios de Gal4p são estáveis quando separados, mas a ativação da RNA-polimerase exige a interação com o domínio de ativação que, por sua vez, necessita de posicionamento próximo do domínio de ligação do DNA. Assim, os domínios devem ser reunidos para funcionar corretamente.

Na **análise de duplo-híbrido em leveduras**, as regiões codificadoras de proteínas gênicas analisadas são fusionadas ao gene da levedura tanto ao domínio de ligação do DNA quanto ao domínio de ativação de Gal4p, e os genes resultantes expressam várias proteínas de fusão. (**Figura 9-21**). Se uma proteína fusionada ao domínio de ligação do DNA interage com uma proteína fusionada ao domínio de ativação, a transcrição é ativada. O gene repórter transcrito por essa ativação é geralmente o que produz uma proteína necessária para o cultivo ou uma enzima que catalisa uma reação com um produto colorido. Desse modo, quando cultivadas em meio adequado, as células que contêm um par de proteínas em interação são facilmente distinguidas das células que não o têm. Uma biblioteca pode ser configurada com uma linhagem de levedura específica, na qual cada célula na biblioteca tem um gene fusionado com o gene do domínio de ligação do DNA de Gal4p, e muitos desses genes estão representados na

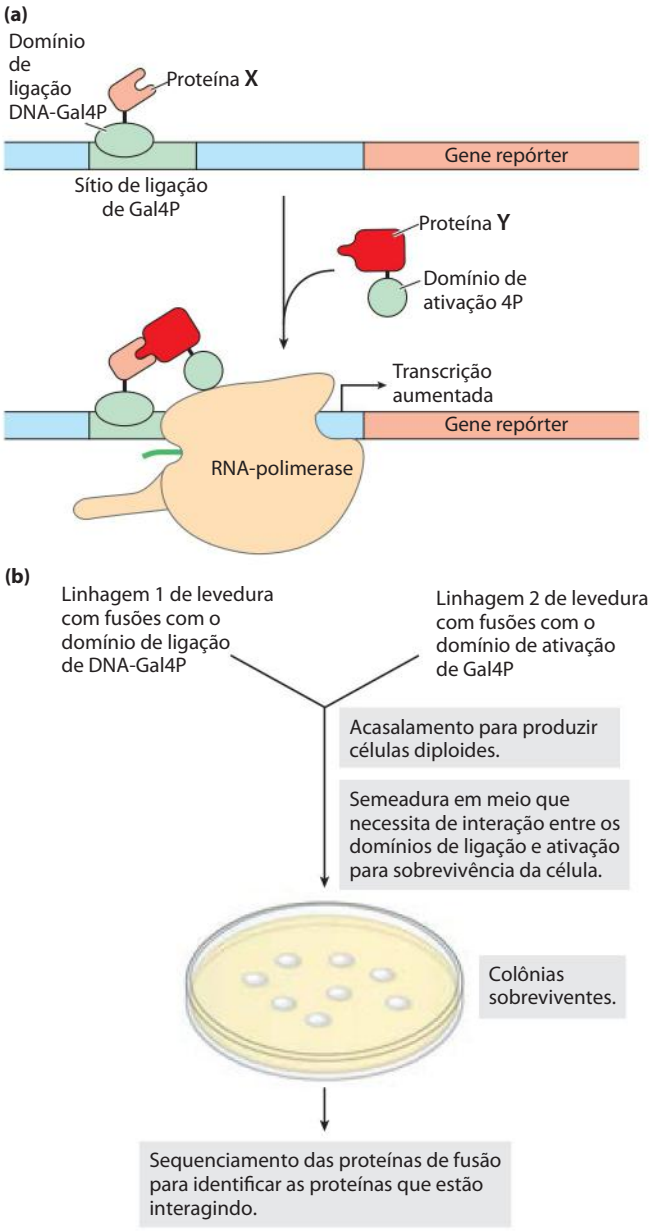


FIGURA 9-21 Análise de duplo-híbrido em leveduras. (a) O objetivo é reunir o domínio de ligação do DNA e o domínio de ativação da proteína de levedura Gal4 (Gal4p) pela interação de duas proteínas, X e Y, à qual um dos domínios se funde. Esta interação é acompanhada pela expressão de um gene repórter. **(b)** As duas fusões de genes são criadas em linhagens de leveduras separadas, sendo então colocadas para reproduzir. A mistura que sofre reprodução é semeada em um meio em que as leveduras não podem sobreviver a menos que o gene repórter seja expresso. Assim, todas as colônias sobreviventes têm proteínas de fusão interagindo. O sequenciamento das proteínas de fusão nas colônias sobreviventes revela quais proteínas estão interagindo.

biblioteca. Em uma segunda linhagem de leveduras, um gene de interesse é fusionado com o gene para o domínio de ativação de Gal4p. As linhagens de leveduras são reproduzidas e as células diploides individuais são cultivadas em colônias. As únicas células que crescem no meio seletivo e produzem a cor adequada são aquelas em que o gene de interesse está ligado a um parceiro, permitindo a transcrição do gene repórter. Isso permite a triagem em larga escala de proteínas celulares que interagem com a proteína-alvo. A proteína de interação, fusionada com o domínio de ligação do DNA de Gal4p, presente em uma determinada colônia selecionada, pode ser rapidamente identificada por sequenciamento de DNA do gene da proteína de fusão. Alguns resultados falsos positivos ocorrem devido à formação de complexos multiproteicos.

Essas técnicas para a determinação da localização celular e interações moleculares fornecem pistas importantes para a função da proteína. Entretanto, elas não substituem a bioquímica clássica. Simplesmente fornecem aos pesquisadores um acesso rápido a novos e importantes problemas biológicos. Quando combinadas às ferramentas da bioquímica e biologia molecular em evolução simultânea, as técnicas descritas aqui aceleram a descoberta não apenas de novas proteínas, mas de novos processos e mecanismos biológicos.

Microarranjos de DNA revelam padrões de expressão de RNA e outras informações

Os principais refinamentos posteriores da tecnologia de bibliotecas de DNA, a PCR e a hibridização se uniram no desenvolvimento de **microarranjos de DNA** que permitem a triagem rápida e simultânea de muitos milhares de genes. Segmentos de DNA de genes de sequências conhecidas, com algumas dezenas de centenas de pares de bases de comprimento, são amplificados por PCR. Dispositivos robóticos então depositam quantidades precisas de nanolitros de soluções de DNA sobre uma superfície sólida de poucos centímetros quadrados, em um arranjo predefinido, com cada um dos milhares de pontos contendo sequências derivadas de um determinado gene. Uma estratégia alternativa e cada vez mais comum para a síntese de DNA diretamente sobre uma superfície sólida utiliza a fotolitografia (**Figura 9-22**). O arranjo resultante, muitas vezes chamado de um *chip*, pode incluir sequências derivadas de cada gene de um genoma

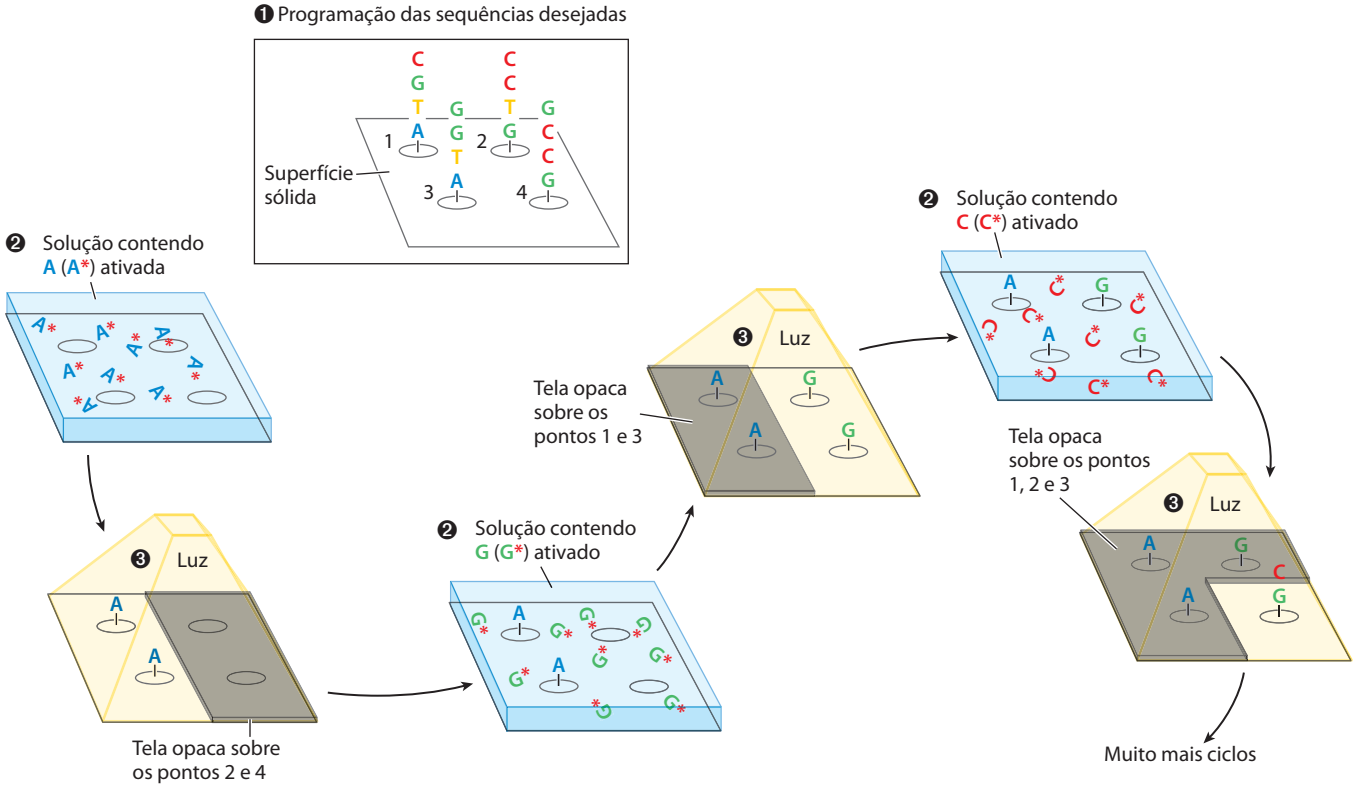
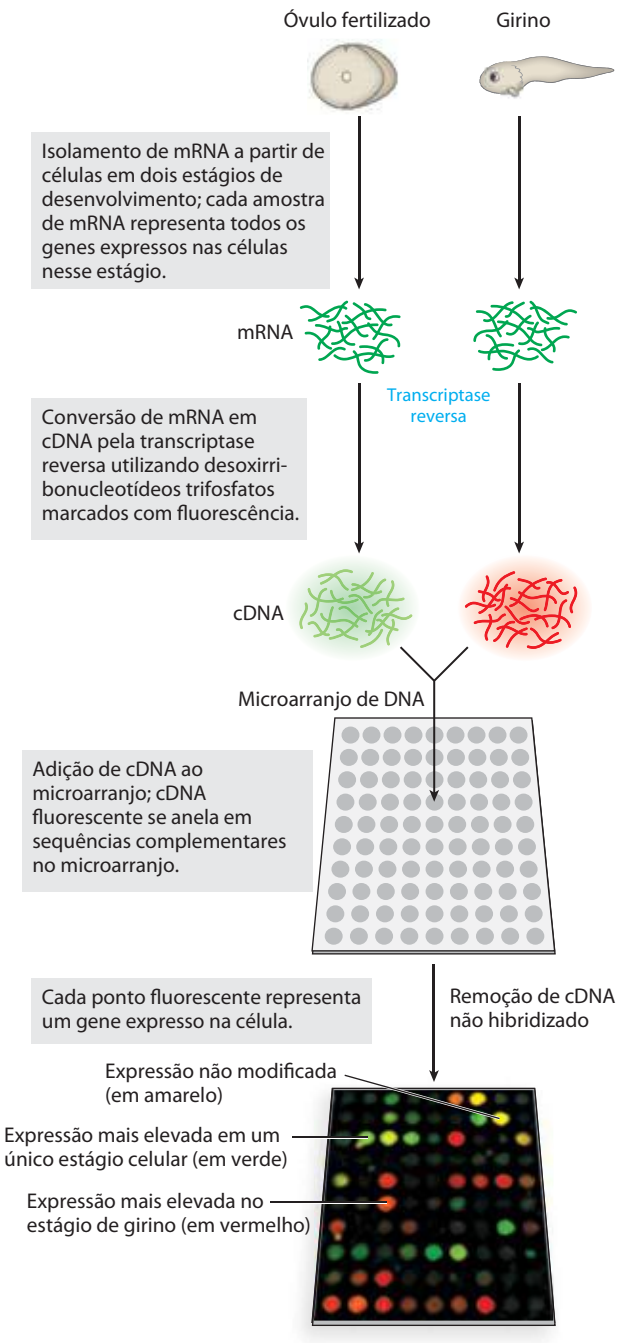


FIGURA 9-22 Fotolitografia para criar um microarranjo de DNA. ❶ Um computador é programado com as sequências de oligonucleotídeos desejadas. ❷ Os grupos reativos, ligados a uma superfície sólida, são inicialmente inativados pelos grupos de bloqueio fotoativos, que podem ser removidos por um *flash* de luz. Uma tela opaca bloqueia a luz de algumas áreas da superfície, impedindo sua ativação. Outras áreas ou "pontos" são expostos. ❸ Uma solução contendo um nucleotídeo ativado (p. ex., A*) é lavada sobre os pontos. O grupo 5′-hidroxila do nucleotídeo é bloqueado para impedir reações indesejadas, e o nucleotídeo se liga aos grupos de superfície em pontos apropriados sobre seu grupo 3′-hidroxila. A superfície é lavada sucessivamente com soluções contendo cada nucleotídeo ativado remanescente (G*, C*, T*). Os grupos bloqueadores 5′ em cada nucleotídeo limitam as reações à adição de um nucleotídeo por vez, e esses grupos também podem ser removidos pela luz. Quando cada ponto tem um nucleotídeo, um segundo nucleotídeo pode ser adicionado para ampliar o nucleotídeo nascente em cada ponto, utilizando telas e luz para assegurar que serão adicionados os nucleotídeos em cada ponto da sequência correta. Este processo continua até que as sequências necessárias sejam construídas em cada um dos milhares de pontos no microarranjo do DNA.

bacteriano ou de levedura, ou de famílias de genes selecionados a partir de um genoma maior. Quando construído, o microarranjo pode ser sondado com mRNA ou cDNA a partir de um determinado tipo celular ou cultura de células para identificar os genes que são expressos nessas células.

Um microarranjo pode fornecer um panorama de todos os genes de um organismo, informando ao pesquisador sobre os genes expressos em determinada fase do desenvolvimento de um organismo ou em um conjunto particular de condições ambientais. Por exemplo, o complemento total de mRNA pode ser isolado a partir de células em duas fases diferentes de desenvolvimento e convertido a cDNA com transcriptase reversa. Desoxirribonucleotídeos marcados por fluorescência podem ser utilizados para tornar uma amostra de cDNA fluorescente vermelha e outra verde (**Figura 9-23**). O cDNA das duas amostras é misturado e utilizado para fazer uma sonda no microarranjo. Cada cDNA se anela em um único ponto no microarranjo, correspondendo ao gene que codifica o mRNA, que deu origem ao cDNA. Pontos que apresentam fluorescência verde representam genes que produzem mRNA em níveis mais elevados em um estágio de desenvolvimento; aqueles que apresentam fluorescência vermelha representam genes expressos em níveis mais elevados em outra fase. Se um gene produz mRNA em igual abundância em ambas as fases de desenvolvimento, o ponto correspondente apresenta fluorescência amarela. Utilizando uma mistura de duas amostras para medir a abundância de sequências relativa e não absoluta, o método corrige as variações na quantidade de DNA originalmente depositado em cada ponto do gráfico, bem como outras possíveis inconsistências entre pontos no microarranjo. Os pontos que apresentam fluorescência fornecem um panorama de todos os genes expressos nas células no momento em que foram colhidos – expressão gênica examinada em uma escala do genoma total. Para um gene de função desconhecida, o tempo e as circunstâncias da sua expressão podem fornecer pistas importantes sobre o seu papel na célula. Essas tecnologias também revelam a expressão de muitos tipos de RNA especializados, tais como microRNA (miRNA; Capítulo 26) e RNA de interferência pequenos (siRNA; Capítulo 28).



Um exemplo dessa técnica é ilustrado na **Figura 9-24**, que mostra os resultados impressionantes produzidos pelos experimentos de microarranjos. Os segmentos de cada um dos cerca de 6.500 genes no genoma de leveduras completamente sequenciado foram amplificados separadamente por PCR, e cada segmento foi depositado em um padrão definido para criar o microarranjo. Em certo sentido, esse arranjo fornece uma visão instantânea do panorama do funcionamento da totalidade do genoma da levedura, em um conjunto de condições.

RESUMO 9.2 Utilização de métodos com base no DNA para a compreensão das funções das proteínas

- As proteínas são estudadas no nível da função fenotípica, celular ou molecular.
- As bibliotecas de DNA são um prelúdio para muitos tipos de investigações que produzem informação sobre a função de proteínas.
- Fusionando um gene de interesse com os genes que codificam a proteína fluorescente verde ou marcadores de epítopos, os pesquisadores podem visualizar a localização celular do produto gênico, tanto diretamente quanto por meio de imunofluorescência.
- As interações de uma proteína com outras proteínas ou com o RNA podem ser investigadas por meio de marca-

FIGURA 9-23 Experimento com microarranjo de DNA. Um microarranjo pode ser preparado a partir de qualquer sequência conhecida de DNA, de qualquer fonte. Quando o DNA se liga a um suporte sólido, outros ácidos nucleicos marcados com fluorescência podem servir como sondas para o microarranjo. Aqui, amostras de mRNA foi retiradas de células de um sapo em dois estágios diferentes de desenvolvimento.

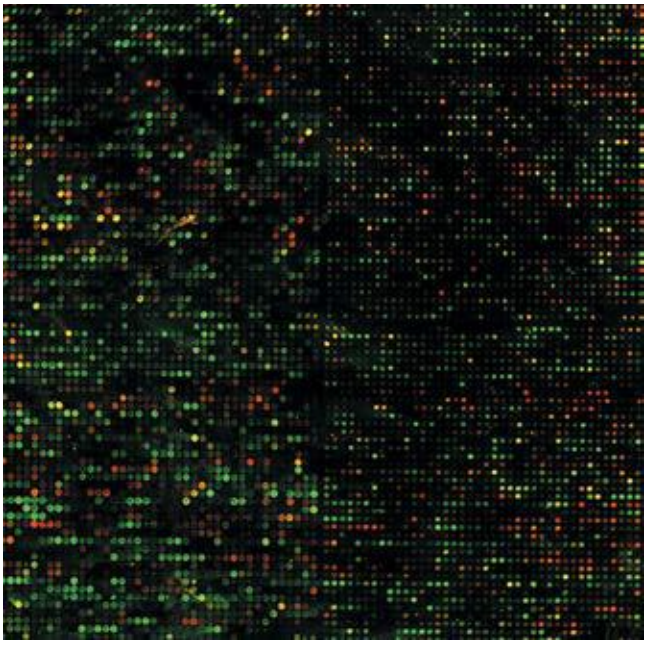


FIGURA 9-24 Imagem ampliada de um microarranjo de DNA. Cada ponto brilhante contém o DNA dos cerca de 6.500 genes do genoma de levedura (*S. cerevisiae*), com cada gene representado no arranjo. As sondas do microarranjo são os ácidos nucleicos marcados com fluorescência derivados de mRNA obtidos nas células em crescimento nos meios de cultura (em verde) e 5 horas após o início da formação dos esporos (em vermelho). Os pontos verdes representam os genes expressos em níveis elevados durante o crescimento; os pontos vermelhos, os genes expressos em níveis elevados durante a esporulação. Os pontos amarelos representam os genes que não mudam seu nível de expressão durante a esporulação. Esta imagem está aumentada; na verdade, o microarranjo mede apenas 1,8 × 1,8 cm.

dores de epítopos e imunoprecipitação ou cromatografia de afinidade. A análise de duplo-híbrido de levedura fornece sondas de interações moleculares *in vivo*.

- Os microarranjos podem revelar padrões de expressão gênica que se alteram em resposta a estímulos, estágio de desenvolvimento ou condições celulares.

9.3 Genômica e história da humanidade

Os métodos de sequenciamento de DNA descritos no Capítulo 8 levaram às primeiras sequências genômicas completas de espécies bacterianas na década de 1990. Duas sequências do genoma humano completo surgiram em 2001. Uma delas foi gerada por um esforço de financiamento público, liderado inicialmente por James Watson e posteriormente por Francis Collins. Em paralelo, esforços privados foram liderados por Craig Venter. Essas conquistas se refletiram em mais de uma década de intensos esforços coordenados em dezenas de laboratórios ao redor do mundo e foram apenas o começo. Os bancos de dados de sequências de DNA contêm agora as sequências genômicas completas de milhares de organismos de todos os tipos. Só agora esses vastos recursos de informação começam a ser compilados de modo eficaz.

Francis S. Collins

J. Craig Venter

O sequenciamento genômico é auxiliado por novas gerações de métodos de sequenciamento de DNA

As tecnologias de sequenciamento de DNA continuam a evoluir. Um genoma humano completo pode agora ser sequenciado em um ou dois dias, e um genoma bacteriano, em poucas horas. O dia em que a sequência genômica de um indivíduo se tornar parte rotineira do registro médico de uma pessoa está próximo (Quadro 9-2). Esses avanços se tornaram possíveis com métodos chamados de sequenciamento de última geração, ou "nextgen". Às vezes, a estratégia de sequenciamento é semelhante ou bastante diferente da utilizada no método de Sanger descrito no Capítulo 8. As inovações permitiram uma miniaturização do procedimento, um grande aumento na escala e uma diminuição correspondente no custo.

Uma sequência genômica é gerada em várias etapas. Primeiro, as sequências genômicas são quebradas em locais aleatórios por cisalhamento para gerar fragmentos de algumas centenas de pares de bases de comprimento. Os oligonucleotídeos sintéticos são ligados às extremidades de todos os fragmentos, fornecendo um ponto de referência conhecido em cada molécula de DNA. Os fragmentos individuais são então imobilizados em uma superfície sólida, e cada um é amplificado utilizando-se a PCR (Figura 9-12). A superfície sólida é parte de um canal que permite que as soluções líquidas possam fluir sobre as amostras. O resultado é uma superfície sólida de apenas alguns centímetros, com milhões de grupamentos de DNA ligados, cada grupo contendo múltiplas cópias de uma única sequência de DNA derivada de um ou de outro fragmento de DNA genômico aleatório. A eficiência é decorrente do sequenciamento de todos esses milhões de grupamentos, ao mesmo tempo, com os dados de cada grupo capturados e armazenados em um computador.

Dois sequenciadores de última geração amplamente utilizados empregam diferentes estratégias para realizar as reações de sequenciamento. Um deles, o sequenciamento 454 (os números se referem a um código utilizado na fase de desenvolvimento da tecnologia e não têm significado científico), utiliza uma estratégia chamada de pirossequenciamento, na qual a adição de nucleotídeos é detectada com *flashes* de luz (**Figura 9-25**). Os quatro desoxinucleosídeos trifosfato (inalterados) são pulsados para a superfície de reação um de cada vez, em uma sequência de repetição. A solução de nucleotídeos fica retida na superfície apenas o tempo suficiente para que a DNA-polimerase possa adicionar esses nucleotídeos em qualquer grupo em que eles sejam complementares à próxima base-molde na sequência. O excesso de nucleotídeos é destruído rapidamente pela en-

QUADRO 9-2 MEDICINA Medicina genômica personalizada

Os gêmeos californianos Noah e Alexis Beery nasceram com sintomas de paralisia cerebral. Os tratamentos não surtiram efeito. Insatisfeitos com o diagnóstico e o tratamento, os pais dos gêmeos, Joe e Retta, os levaram, com a idade de 5 anos, a um especialista em Michigan, que os diagnosticou com uma rara condição genética, chamada de distonia responsiva à DOPA. Um regime de tratamento suprimiu com sucesso os sintomas e permitiu que os gêmeos levassem uma vida normal. Entretanto, aos 12 anos, Alexis desenvolveu tosse severa e dificuldades para respirar que novamente pareciam ameaçar a vida da menina. Em um episódio, os paramédicos tiveram que reanimá-la duas vezes. Os sintomas não pareciam estar relacionados à distonia. Seria Noah o próximo? Frustrados e profundamente preocupados, os pais dos gêmeos procuraram o sequenciamento completo do genoma tanto de Noah quanto de Alexis. Essa decisão aparentemente incomum foi natural para a família Beery. Joe era o presidente da Life Technologies, desenvolvedores de tecnologias de sequenciamentos em uso por muitos grandes centros de sequenciamento de DNA. Os casos de Noah e Alexis foram assumidos por Matthew Bainbridge e sua equipe no Baylor College of Medicine Human Genome Sequencing Center, em Houston, Texas. Os resultados mostraram-se decisivos. Os gêmeos apresentavam mutações em seus genomas que produziram não somente uma deficiência na DOPA, mas também uma deficiência potencial na produção do hormônio serotonina. Um pequeno ajuste na terapia de Alexis lhe proporcionou o fim dos sintomas que colocavam sua vida em risco, e o mesmo tratamento foi oferecido ao seu irmão. Os dois irmãos agora levam uma vida normal.

O primeiro projeto de sequenciamento do genoma humano foi concluído em 2001, após 12 anos, a um custo de US$ 3 bilhões. Esse custo caiu (Figura Q-1), e hoje genomas humanos completos são comuns. O objetivo de 1.000 dólares por genoma por pessoa está no horizonte e promete tornar essa tecnologia amplamente disponível. Como a maioria das alterações genômicas que afetam a saúde humana está em genes codificadores de proteínas (suposição que pode ser contestada nos próximos anos), uma alternativa mais barata é simplesmente sequenciar o 1% do genoma que representa as regiões codificadoras (éxons) de genes, ou o **exoma**.

Essa primeira sequência do genoma humano veio de um genoma haploide, derivado de um amálgama de DNA de vários seres humanos diferentes. A conclusão de uma sequência de alta qualidade, em 2004, estabeleceu essa sequência como o genoma humano de referência. Sequências do genoma humano concluídas posteriormente, muitas de genomas diploides individuais, demonstraram o quanto existe de variação genética individual. Em relação à sequência de referência, um ser humano comum tem cerca de 3,5 milhões de SNP (cerca de 60% desses são heterozigotos, presentes apenas em um dos dois cromossomos) e algumas outras centenas de milhares de diferenças na forma de pequenas inserções e deleções e alterações no número de cópias repetidas. Apenas uma pequena porção (5.000 a 10.000) dos SNP afeta as sequências de aminoácidos das proteínas codificadas pelos genes.

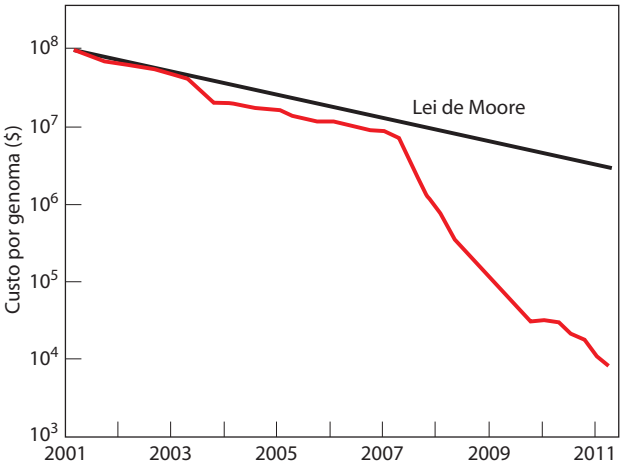
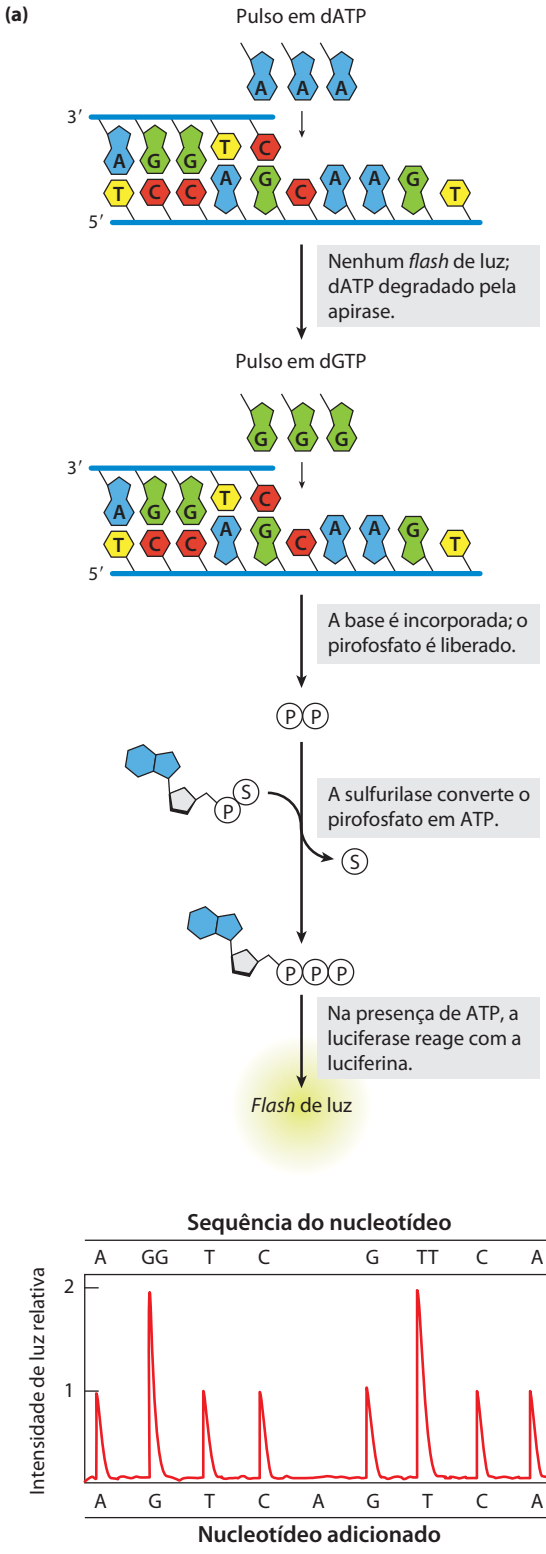


FIGURA Q-1 Desde janeiro de 2008, o custo do sequenciamento do genoma humano tem diminuído mais rápido do que o declínio projetado no custo do processamento de dados pelo computador (Lei de Moore).

Essa complexidade assegura que, pelo menos no curto prazo, o diagnóstico bem-sucedido de uma condição pelo sequenciamento de todo o genoma será uma exceção, e não a regra. Entretanto, poucas disciplinas avançam tão rapidamente quanto a genômica humana. O número de casos de sucesso aumenta rapidamente à medida que a tecnologia se torna mais amplamente disponível e se aprimora a capacidade das análises genômicas para reconhecer alterações genéticas causadoras de doenças.

zima apirase antes que o próximo nucleotídeo pulse. Quando um nucleotídeo específico é adicionado com sucesso às cadeias de um agrupamento, o pirofosfato é liberado como um supbroduto. Outra enzima na solução que banha a superfície é a sulfurilase, que converte o pirofosfato em ATP.

O aparecimento do ATP fornece por fim o sinal de que um nucleotídeo foi adicionado ao DNA. Também estão presentes no meio a enzima denominada luciferase e uma molécula do substrato, a luciferina (a luciferase é a enzima que gera o *flash* de luz produzido por vaga-lumes, ver Quadro 13-1). Quando o ATP é gerado, a luciferase catalisa uma reação com a luciferina que resulta em um pequeno *flash* de luz. Quando muitos *flashes* pequenos ocorrem em um grupo, a luz emitida pode ser gravada em uma imagem capturada. Por exemplo, quando o dCTP é adicionado à solução, os *flashes* irão ocorrer apenas nos grupos onde G está presente no molde e C é o próximo nucleotídeo a ser adicionado à cadeia de DNA em formação. Se houver uma fita de 2, 3 ou 4



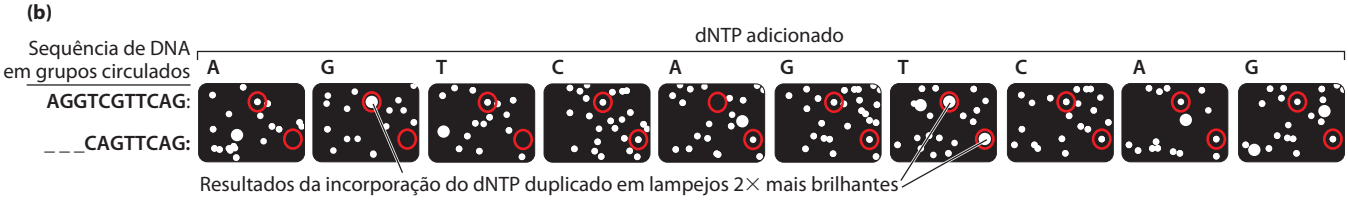


FIGURA 9-25 **Pirossequenciamento de última geração.** **(a)** O pirossequenciamento utiliza as enzimas sulfurilase (ver Figura 22-15) e luciferase (ver Quadro 13-1) para detectar a adição de nucleotídeos com *flashes* de luz. O gráfico mostra as intensidades de luz observadas durante ciclos sucessivos de sequenciamento para um segmento de DNA imobilizado em um ponto específico de uma placa de microtitulação e (no topo) a sequência de nucleotídeos do DNA derivada deles. **(b)** Uma imagem de uma parte muito pequena de um ciclo do sequenciamento 454. Cada segmento individual de DNA a ser sequenciado é ligado a um micrograânulo de captura de DNA, em seguida amplificado no grânulo por PCR. Cada grânulo é imerso em uma emulsão e colocado em um micropoço (~29 μm) em uma placa de microtitulação. A reação de luciferina e do ATP com a luciferase produz *flashes* de luz quando um nucleotídeo é adicionado a um grupo de DNA específico em um determinado poço. Os círculos representam o mesmo grupo após múltiplos ciclos. Nesse caso, a leitura do ciclo de cima (ou de baixo) da esquerda para a direita ao longo de cada seta fornece a sequência para esse grupo.

resíduos G no molde, um número semelhante de resíduos C serão adicionados à fita em formação em um ciclo. Isso será registrado como uma amplitude de *flash* nesse grupo 2, 3 ou 4 vezes maior do que quando apenas um resíduo C é adicionado. Do mesmo modo, quando é adicionado dGTP, ocorrem *flashes* em um conjunto diferente de grupos, marcando-os como grupos em que G é o próximo nucleotídeo adicionado à sequência. O comprimento do DNA que pode ser sequenciado com confiança em um grupo único por esse método – muitas vezes chamado de comprimento lido ou "*read*" – normalmente é de 400 a 500 nucleotídeos, como no início de 2012, e está sendo constantemente aperfeiçoado.

O sequenciador Illumina é o método alternativo e utiliza uma técnica conhecida como sequenciamento de terminação reversível (**Figura 9-26**). Um oligonucleotídeo iniciador de sequenciamento especial é adicionado, que é complementar aos oligonucleotídeos de sequência conhecida que foram ligados às extremidades dos fragmentos de DNA em cada grupo (como descrito anteriormente). Além disso, nucleotídeos terminadores marcados com fluorescência e DNA-polimerase são adicionados. A polimerase adiciona o nucleotídeo adequado às fitas em cada grupo, cada tipo de nucleotídeo (A, T, G ou C) transportando um marcador fluorescente diferente. Esses nucleotídeos terminadores têm grupos de bloqueio ligados às estremidades 3′, permitindo apenas a adição de um nucleotídeo a cada fita. Em seguida, raios lasers excitam todos os marcadores fluorescentes, e uma imagem de toda a superfície revela a cor (e, portanto, a identidade da base) adicionada a cada grupo. O marcador fluorescente e os grupos de bloqueio são então removidos química ou fotoliticamente, em preparação para a adição de um novo nucleotídeo para cada grupo. Os procedimentos de sequenciamento são realizados em etapas. Os comprimentos lidos são mais curtos para esse método, normalmente de 100 a 200 nucleotídeos por grupo.

Essas tecnologias são manifestações modernas de um método de sequenciamento genômico às vezes chamado de sequenciamento *shotgun*. Muitas cópias do DNA genômico

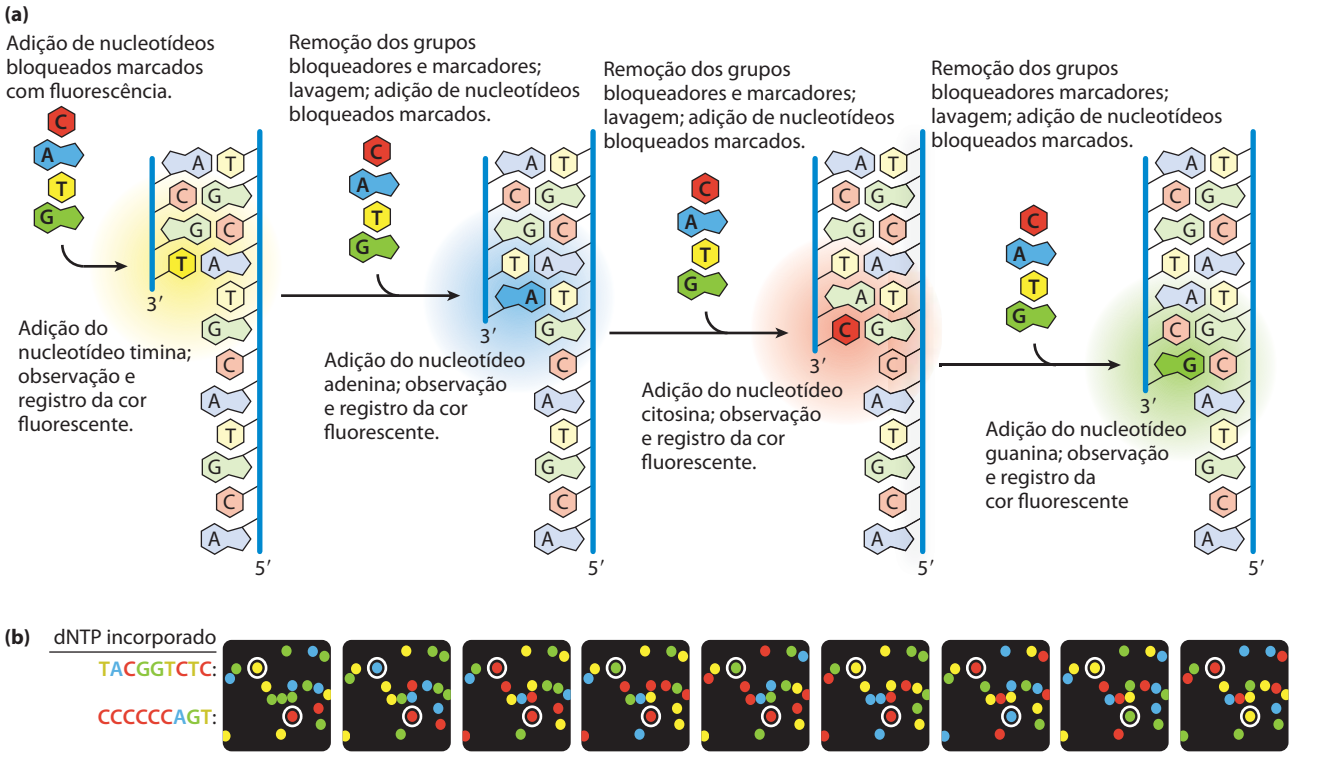


FIGURA 9-26 **Sequenciamento de terminação reversível de última geração.** **(a)** O método de terminação reversível utiliza marcadores fluorescentes para identificar nucleotídeos. Grupos de bloqueio em cada nucleotídeo marcado com fluorescência evitam a adição de múltiplos nucleotídeos por ciclo. **(b)** Seis ciclos sucessivos a partir de uma parte muito pequena de um sequenciamento Illumina. Cada ponto colorido representa a localização de um oligonucleotídeo de DNA fixo na superfície da célula de fluxo. Os grupos circulados representam o mesmo ponto na superfície após ciclos sucessivos e fornecem as sequências indicadas. Os dados são automaticamente gravados e digitalmente analisados. **(c)** Célula de fluxo típica utilizada para um sequenciador de última geração. Milhões de fragmentos de DNA podem ser sequenciados simultaneamente em cada um dos oito canais.

são cortadas para gerar cada conjunto de fragmentos. Assim, um determinado segmento curto do genoma pode estar presente em dezenas ou mesmo centenas de diferentes grupos sequenciados. Entretanto, não há qualquer marcação sobre um fragmento individual para revelar a procedência do genoma. Traduzir as sequências desses milhões de fragmentos em uma sequência genômica requer o alinhamento computadorizado de sequências de fragmentos que se sobrepõem (**Figura 9-27**). As sobreposições permitem rastrear a sequência ao longo de um cromossomo no computador, de um fragmento para outro. Isso permite a montagem de sequências contíguas chamadas de **contigs**. Em um exercício de sequenciamento genômico bem-sucedido, muitos contigs podem se estender por milhões de pares de bases. São necessárias estratégias especiais para preencher as lacunas inevitáveis e para lidar com as sequências repetitivas.

O genoma humano contém genes e muitos outros tipos de sequências

Métodos novos e mais eficientes para o sequenciamento de DNA produziram uma explosão de novas sequências genômicas. Elas são armazenadas e disponibilizadas em vários bancos de dados acessíveis publicamente. Uma boa abordagem para esses recursos pode ser encontrada no National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov). A análise dessa riqueza crescente de informações genômicas está em curso. O que revela o genoma humano e sua comparação com o de outros organismos?

Em alguns aspectos, o ser humano não é tão complicado como se imaginava. A estimativa de décadas atrás de que os seres humanos tivessem cerca de 100.000 genes nos aproximadamente $3{,}2 \times 10^9$ pares de bases do genoma humano foi suplantada pela descoberta de que existem apenas cerca de 25.000 genes codificadores de proteínas – menos do que o dobro do número em uma mosca-das-frutas (13,6 mil), não muito mais do que em um nematódeo (20.000) e menos do que em uma planta de arroz (38.000).

Em outros aspectos, entretanto, o ser humano é mais complexo do que se pensava anteriormente. O estudo da estrutura dos cromossomos eucarióticos e das sequências do genoma revelou que muitos, se não a maioria dos genes eucarióticos, contêm um ou mais segmentos de DNA intervenientes que não codificam a sequência de aminoá-

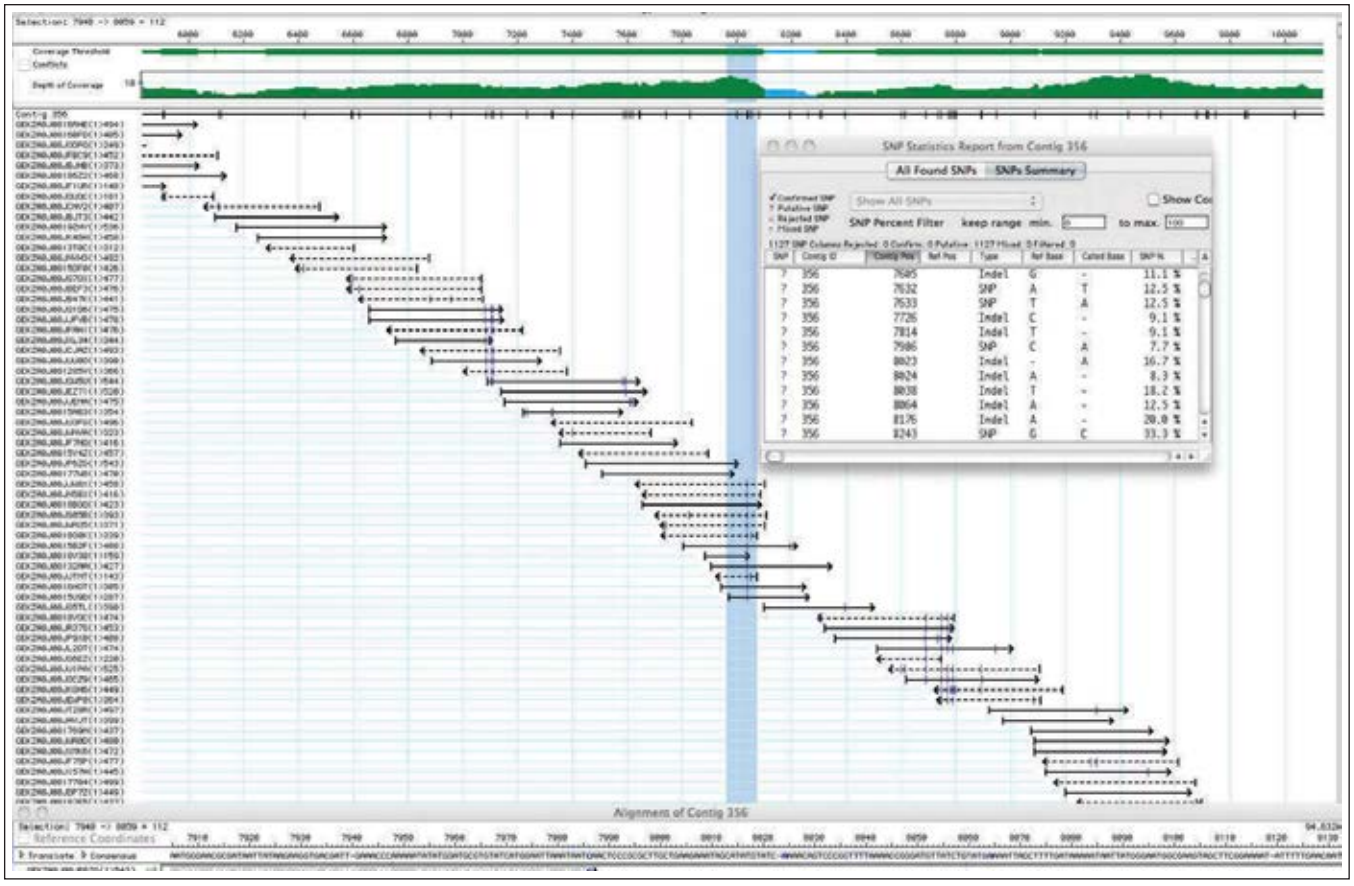


FIGURA 9-27 **Montagem da sequência.** Em uma sequência genômica, cada par de bases do genoma costuma ser representado em vários, frequentemente dezenas, dos fragmentos sequenciados, chamados de *reads*. É mostrada uma pequena parte da sequência de uma nova variante de espécie de *E. coli*, com os *reads* gerados por um sequenciador 454. Os números no topo representam as posições dos pares de bases genômicos, em relação a um "0" definido arbitrariamente. Todas as sequências são provenientes de um contig longo denominado 356. Os *reads* são representados por setas horizontais, com identificadores atribuídos por computador listados para cada um à esquerda. Segmentos de fita de DNA são sequenciados aleatoriamente, com sequências obtidas de uma fita (5′→3′, da esquerda para a direita) representadas por setas sólidas e sequências obtidas de outra fita (5′→3′, da direita para a esquerda) representadas por linhas pontilhadas. As últimas sequências são automaticamente relatadas como seus complementos quando são fundidas com o conjunto de dados em geral. O limiar de cobertura na parte superior é uma medida da qualidade da sequência. A barra verde maior indica as sequências que foram obtidas em número de vezes suficientes para gerar alta fidelidade nos resultados. A profundidade da linha de cobertura indica quantas vezes um determinado par de bases aparece em um *read* sequenciado. A barra azul vertical denota uma parte da sequência destacada na linha de sequência na parte inferior da figura. O relatório estatístico de SNP (no detalhe) é uma lista de posições onde polimorfismos de nucleotídeos únicos parecem estar presentes em alguns dos *reads*. Esses supostos SNP são muitas vezes verificados por sequenciamento adicional. Eles são indicados nos *reads* por marcas verticais finas azuis nas linhas horizontais de cada *read*.

cidos do produto polipeptídico. Essas inserções não traduzidas interrompem, por outro lado, a relação colinear entre a sequência de nucleotídeos do gene e a sequência de aminoácidos do polipeptídeo codificado. Esses segmentos de DNA não traduzidos são chamados de **sequências intervenientes** ou **íntrons**, e os segmentos codificadores são chamados de **éxons** (**Figura 9-28**). Poucos genes bacterianos contêm íntrons. Os íntrons são removidos do transcrito de RNA primário e os éxons são unidos para gerar um transcrito que pode ser traduzido juntamente em um produto proteico (ver Capítulo 26). Um éxon muitas vezes (mas nem sempre) codifica um único domínio de uma proteína maior com vários domínios. Modos alternativos de expressão gênica e *splicing* do RNA permitem a produção de várias combinações de éxons, que levam à produção de mais de uma proteína a partir de um único gene. Os seres humanos compartilham muitos tipos de domínios de pro-

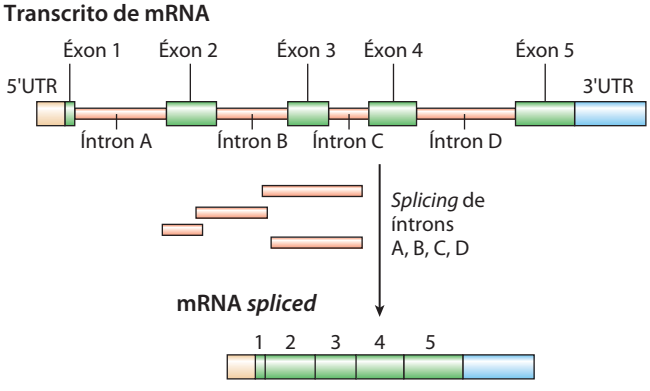


FIGURA 9-28 **Íntrons e éxons.** Esse transcrito de gene contém cinco éxons e quatro íntrons, ao longo das regiões 5′e 3′ não traduzidas (5′URT e 3′URT). O *splicing* remove os íntrons para criar um produto de mRNA para a tradução na proteína.

teínas com plantas, vermes e moscas, mas esses domínios são utilizados em arranjos mais complexos, gerando, assim, proteínas mais complexas.

Nos mamíferos e em alguns outros eucariotos, um gene normal tem uma proporção muito maior de DNA íntron do que DNA éxon; na maioria dos casos, a função dos íntrons não é clara. Apenas cerca de 1,5% do DNA humano é "codificador" ou DNA éxon, transportando informação para os produtos proteicos (**Figura 9-29a**). Entretanto, quando os íntrons muito maiores são incluídos na contagem, cerca de 30% do genoma humano consiste nesses genes. Vários esforços estão em curso para categorizar os genes codificadores de proteínas por função (Figura 9-29b).

A relativa escassez de genes no genoma humano deixa uma grande quantidade de DNA inexplicado. Grande parte do DNA no gene está na forma de sequências repetidas de vários tipos. Talvez o mais surpreendente seja que cerca de metade do genoma humano é constituída por sequências moderadamente repetidas derivadas de elementos transponíveis – segmentos de DNA que vão de algumas centenas a vários milhares de pares de bases de comprimento, que podem se mover de um local para outro no genoma. Originalmente descobertos no milho por Barbara McClintock, os elementos transponíveis, ou **transposons**, são uma espécie de parasita molecular. Eles fazem sua casa nos genomas de essencialmente todos os organismos de modo eficiente e em grande parte passivo. Muitos transposons contêm genes que codificam proteínas que catalisam o próprio processo de transposição, como detalhado nos Capítulos 25 e 26. Existem múltiplas classes de transposons no genoma humano. Alguns são ativos, movendo-se em frequência baixa, mas a maioria é inativo, relíquias evolutivas alteradas por mutações.

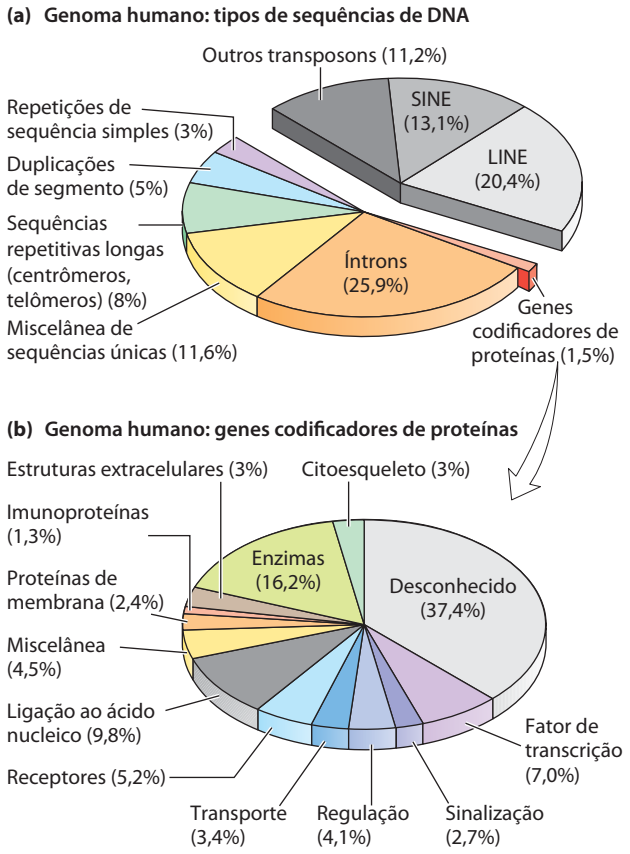


FIGURA 9-29 Panorama do genoma humano. (a) Este gráfico setorial mostra as proporções de vários tipos de sequências no genoma humano. Os transposons, incluindo os SINE e LINE, são descritos nos Capítulo 25 e 26. **(b)** Os aproximadamente 25.000 genes codificadores de proteínas no genoma humano podem ser classificados pelo tipo de proteína que codificam.

Uma vez que os genes codificadores de proteínas (incluindo éxons e íntrons) e transposons sejam levados em conta, talvez restem cerca de 25% do DNA total. A maior parcela disso consiste em sequências únicas encontradas entre genes codificadores de proteínas. Como descrito no Capítulo 26, praticamente todos esses segmentos de DNA são transcritos em RNA em pelo menos algumas células humanas. Novas classes de RNA funcionais – codificadas por genes cuja existência era previamente insuspeita – estão sendo descobertas em um ritmo acelerado. Muitos genes codificadores de RNA funcionais são difíceis de identificar pelos métodos automatizados, em particular quando os produtos de RNA não foram caracterizados. Entretanto, os genes codificadores de RNA são claramente uma característica marcante dessas regiões genômicas, que de outro modo, seriam desconhecidas.

Outros 3% ou menos do genoma humano são compostos de sequências altamente repetitivas chamadas de **repetições de sequências simples** (**SSR**). Geralmente com menos de 10 pb de comprimento, uma SSR é algumas vezes repetida milhões de vezes por célula e tem uma importância funcional identificável no metabolismo de células humanas. Os exemplos mais importantes de DNA SSR ocorrem em centrômeros e telômeros (ver Capítulo 24). Entretanto, repetições longas de sequências simples também ocorrem por todo o genoma.

O que todas essas informações nos revelam sobre as semelhanças e diferenças entre os seres humanos individuais? Dentro da população humana há milhões de variações de base única, chamadas de **polimorfismos de nucleotídeo único**, **ou SNP** (pronuncia-se "snips"). Cada ser humano difere do próximo, em média, em 1 em cada 1.000 pb. Muitas dessas variações estão na forma de SNP, mas também ocorre uma grande variedade de deleções maiores, inserções e pequenos rearranjos na população humana. A partir dessas diferenças genéticas, muitas vezes sutis, surgem as variações humanas perceptíveis – diferenças na cor do cabelo, estatura, visão, alergias, tamanho do pé e (até certo grau desconhecido) comportamento.

O processo de recombinação genética e a segregação cromossômica durante a meiose tendem a misturar e combinar essas pequenas variações genéticas para que diferentes combinações de genes sejam herdadas (ver Capítulo 25). Quando duas dessas variações genéticas estão em cromossomos diferentes, as variantes específicas que possam ser herdadas por um determinado indivíduo são o resultado do acaso. Se as variantes genéticas estão no mesmo cromossomo, a chance delas serem herdadas em conjunto é uma função inversa da distância entre elas. Grupos de SNP e outras diferenças genéticas próximas no mesmo cromossomo são raramente afetados por recombinação e costumam ser herdados juntos; esses grupos são conhecidos como **haplótipos**. Os haplótipos fornecem marcadores

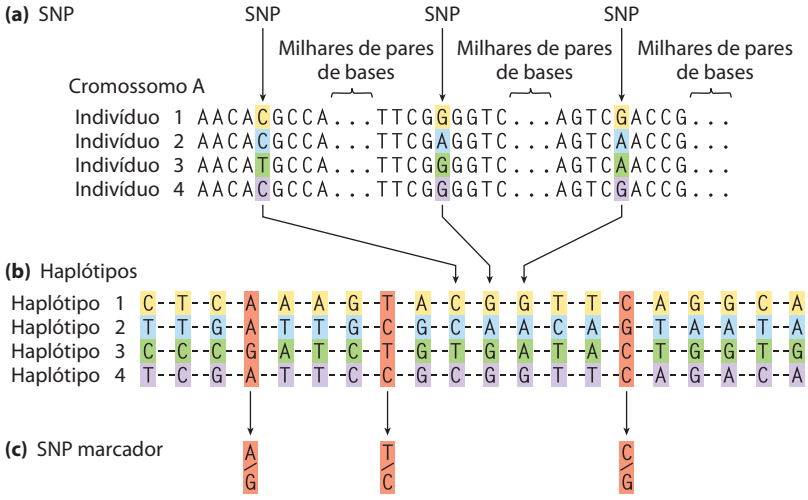


FIGURA 9-30 Identificação de haplótipos. (a) Polimorfismos de nucleotídeo único (SNP), ou posições no genoma humano onde as sequências variam de um indivíduo para outro, são identificadas em amostras genômicas e **(b)** grupos de SNP relativamente próximos uns dos outros em um cromossomo (dentro de algumas dezenas de milhares de pares de bases) são compilados em um haplótipo. Os SNP irão variar na população em geral, assim como nos quatro indivíduos fictícios mostrados nos painéis (a) e (b). Entretanto, os SNP escolhidos para definir um haplótipo são frequentemente os mesmos na maioria dos indivíduos de uma determinada população. **(c)** SNP definidores de haplótipos (SNP marcadores) podem ser utilizados para simplificar o processo de identificação de um haplótipo de um indivíduo (sequenciando 3 SNP definidores, em vez de todos os 20). Se os SNP marcadores mostrados em vermelho claro forem sequenciados, os nucleotídeos A, T e G nas três posições sucessivas do SNP marcador poderiam ser característicos de uma população nativa de um local no norte da Europa, enquanto os nucleotídeos G, T e C nessas três mesmas posições genômicas poderiam ser encontrados em uma população da Ásia. Múltiplos haplótipos deste tipo são utilizados para rastrear migrações humanas pré-históricas (ver Figura 9-35).

convenientes para algumas populações humanas e indivíduos dentro das populações.

Para definir um haplótipo são necessárias várias etapas. Em primeiro lugar, as posições que contêm SNP na população humana são identificadas em amostras de DNA genômico a partir de múltiplos indivíduos (**Figura 9-30a**). Cada SNP pode ser separado do seguinte por muitos milhares de pares de bases. Em segundo lugar, os SNP relativamente próximos em um cromossomo e, portanto, geralmente herdados juntos, são compilados em haplótipos (Figura 9-30b). Cada haplótipo consiste nas bases específicas encontradas nas várias posições do SNP no haplótipo definido. Finalmente, SNP marcadores – um subconjunto de SNP que define todo o haplótipo – são escolhidos para identificar cada haplótipo individualmente (Figura 9-30c). Por meio do sequenciamento de apenas essas posições marcadoras em amostras genômicas de populações humanas, os pesquisadores rapidamente identificam quais haplótipos estão presentes em cada indivíduo. Haplótipos especialmente estáveis existem no genoma mitocondrial (que, sendo herdados pela mãe, nunca sofrem recombinação meiótica) e no cromossomo Y masculino (apenas 3% dos quais são homólogos ao cromossomo X e, portanto, sujeitos a recombinação).

O sequenciamento do genoma dá informações sobre a humanidade

A finalidade principal da maioria dos projetos de sequenciamento do genoma é identificar elementos genéticos conservados de significado funcional, como as sequências conservadas dos éxons, regiões reguladoras e outras características genômicas, como centrômeros e telômeros. Um dos principais objetivos do sequenciamento do genoma humano é a identificação das diferenças entre o genoma humano e dos outros organismos. Embora o genoma humano esteja intimamente relacionado a genomas de outros mamíferos em amplos segmentos de cada cromossomo, as diferenças de uns poucos porcentuais em bilhões de pares de bases adicionam milhões de distinções genéticas. A pesquisa sobre essas diferenças utilizando técnicas genômicas comparativas revela a base molecular de doenças genéticas humanas e ajuda a identificar genes, alterações gênicas e outras características genômicas exclusivamente humanas e, assim, contribuir para definir características humanas como cérebro grande, habilidades linguísticas, a capacidade de produzir ferramentas ou o bipedalismo.

As sequências do genoma de parentes biológicos mais próximos, os chimpanzés e bonobos, oferecem algumas pistas importantes e ilustram o processo comparativo. Os seres humanos e os chimpanzés compartilharam um ancestral comum cerca de 7 milhões de anos atrás. As diferenças genômicas entre as duas espécies são de dois tipos: mudanças de pares de base (SNP) e rearranjos genômicos maiores de muitos tipos. Os SNP em regiões codificadoras de proteínas com frequência resultam em mudanças nos aminoácidos que podem ser utilizadas para construir uma árvore filogenética (**Figura 9-31a**). Os segmentos de cromossomos podem se inverter ao longo do processo evolutivo. Os processos que levam a tais inversões são complicados e raros, mas a linhagem humana tem segmentos longos de DNA invertidos (em relação a outros primatas) em função de processos nos cromossomos 1, 12, 15, 16 e 18. Fusões cromossômicas também podem ocorrer. Na linhagem humana, dois cromossomos encontrados em outras linhagens de primatas se fusionaram para formar o único cromossomo 2 humano (Figura 9-31b).

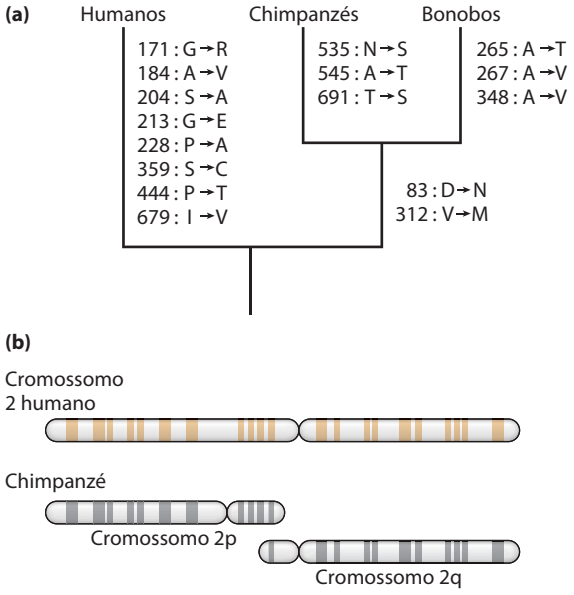


FIGURA 9-31 Alterações genômicas na linhagem humana. (a) Esta árvore evolutiva para primatas é derivada de sequências para o receptor de progesterona, que ajudam a regular muitos eventos na reprodução. O gene que codifica essa proteína sofreu mais alterações evolutivas do que a maioria. Mudanças nos aminoácidos associadas unicamente a humanos, chimpanzés e bonobos são listadas pelo número de resíduos ao lado de cada ramo. **(b)** Os genes nos cromossomos de chimpanzés 2p e 2q são homólogos àqueles no cromossomo 2 humano, indicando que, em algum ponto na linhagem que leva aos humanos, os dois cromossomos se fusionaram em um.

Assim, a linhagem humana tem 23 pares de cromossomos em vez dos 24 pares normais em outros primatas. Quando essa fusão apareceu na linha que conduz aos seres humanos, ela representou uma grande barreira para o cruzamento com outros primatas que não a apresentavam.

Sem considerar os transposons e os grandes rearranjos cromossômicos, os genomas humanos e de chimpanzés publicados diferem em apenas 1,23% no nível de pares de bases (em comparação com a variação de 0,1% de um ser humano para outro). Sem considerar mais variações nas posições onde existe um polimorfismo conhecido tanto na população humana quanto na de chimpanzés (essas pouco provavelmente refletem uma mudança evolutiva que define uma nova espécie), as diferenças aumentam para cerca de 1,06%, ou cerca de 1 em cada 100 pb. Essa pequena porcentagem se traduz em mais de 30 milhões de mudanças de pares de base, algumas das quais afetam a função de proteínas e a regulação gênica. Os rearranjos genômicos que ajudam a distinguir os chimpanzés e os seres humanos incluem 5 milhões de inserções curtas ou deleções envolvendo alguns pares de bases cada, bem como um número substancial de inserções maiores, deleções, inversões ou duplicações que podem envolver muitos milhares de pares de bases. Quando inserções de transposons – importante fonte de variação genômica – são incluídas, as diferenças entre os genomas humano e de chimpanzé aumentam para cerca de 90 milhões de pb, representando outros 3% desses genomas. Com efeito, cada espécie tem segmentos de DNA, constituindo 40 a 45 milhões de pb totalmente exclusivos para esse genoma particular, com inserções cromossômicas maiores, duplicações e outros rearranjos que afetam mais pares de bases do que fazem as mudanças de nucleotídeo único. Portanto, as diferenças genômicas totais entre chimpanzés e humanos equivalem a cerca de 4% de seus genomas.

Classificar as diferenças genômicas relevantes para características exclusivamente humanas é uma tarefa desencorajadora. Se as duas espécies compartilham um ancestral comum, então, assumindo uma taxa de evolução semelhante em ambas as linhagens, metade das mudanças representa mudanças na linhagem dos chimpanzés e a outra metade alterações na linhagem humana. Quando se observa uma diferença, como descrever qual variante era aquela presente no ancestral comum? Uma maneira é comparar ambas as sequências do genoma com as de organismos relacionados mais distantemente, chamados de **grupos externos**. Considere um *locus* X, onde existe uma diferença entre o genoma humano e o do chimpanzé (**Figura 9-32**). A linhagem do orangotango, um grupo externo, divergiu da de chimpanzés e humanos antes do ancestral comum chimpanzé-humano. Se a sequência no *locus* X é idêntica em orangotangos e chimpanzés, essa sequência estava provavelmente presente no ancestral chimpanzé-humano, e a sequência observada em humanos é específica para a linhagem humana. As sequências idênticas em seres humanos e nos orangotangos podem ser eliminadas como candidatas para características genômicas humanas específicas. A importância de comparações com grupos externos intimamente relacionados originou novos esforços para sequenciar o genoma de orangotangos, macacos e muitas outras espécies de primatas.

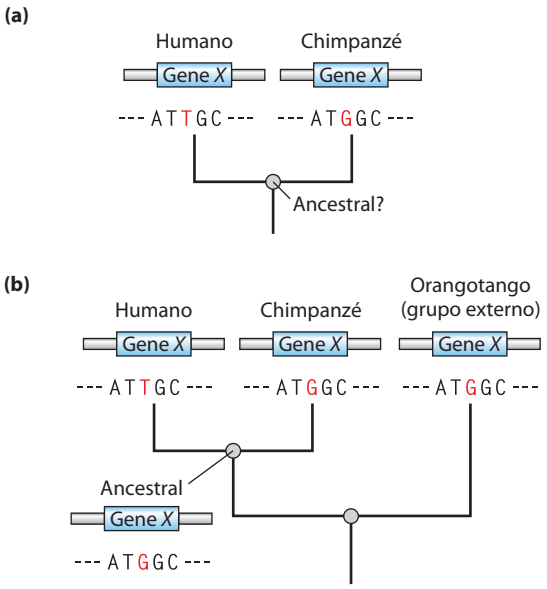


FIGURA 9-32 Determinação das alterações de sequências únicas para uma linhagem ancestral. (a) São comparadas as sequências do mesmo gene hipotético em humanos e chimpanzés. A sequência deste gene em seu último ancestral comum é desconhecida. **(b)** O genoma do orangotango é utilizado como um grupo externo. A sequência do gene do orangotango é idêntica à do gene do chimpanzé. Isso significa que a mutação que leva a diferenças entre humanos e chimpanzés quase certamente ocorreu na linhagem que conduziu aos humanos modernos, e o ancestral comum de humanos e chimpanzés (e orangotangos) tinha a sequência agora encontrada em chimpanzés.

A busca pelas bases genéticas de características humanas especiais, tais como a função cerebral avançada, beneficia-se de duas abordagens complementares. As primeiras pesquisas de regiões genômicas mostram alterações extremas a partir de outros primatas. Essas incluem várias duplicações de genes ou a adição de segmentos genômicos grandes não presentes em outros primatas. A segunda abordagem analisa genes conhecidos envolvidos em condições humanas relevantes. Para a função cerebral, por exemplo, pode-se examinar genes que, quando mutados, contribuem para os transtornos cognitivos ou outros distúrbios mentais.

Notavelmente, as análises da linhagem humana ainda não detectaram um enriquecimento de modificações genéticas nos genes que codificam proteínas envolvidas no desenvolvimento ou no tamanho cerebral. Em primatas, a maioria dos genes que funcionam exclusivamente no cérebro é ainda mais altamente conservada do que os genes que funcionam em outros tecidos. Entretanto, algumas diferenças na expressão gênica são observadas. Quando as mudanças em regiões genômicas relacionadas à regulação gênica são analisadas, os genes envolvidos no desenvolvimento neural e nutrição são desproporcionalmente afetados. Diversos genes codificadores de RNA, alguns com expressão concentrada no cérebro, também mostram evidências de evolução acelerada (**Figura 9-33**). As muitas novas classes de RNA que estão sendo descobertas (ver Capítulo 26) provavelmente irão mudar radicalmente a perspectiva sobre como a evolução altera o funcionamento dos sistemas de seres vivos. É cada vez mais evidente que os genes expressos talvez não sejam tão importantes como quando, onde e quanto deles é expresso.

Comparações do genoma ajudam a localizar genes envolvidos em doenças

O Projeto Genoma Humano tem cumprido o seu potencial para acelerar a descoberta de genes subjacentes a doenças genéticas: mais de 1.600 doenças genéticas humanas já tiveram seus genes específicos mapeados. Com o uso de um método denominado **análise de ligação**, o gene envolvido em uma doença é mapeado em relação aos polimorfismos genéticos bem caracterizados que ocorrem em todo o genoma humano. A busca muitas vezes começa com uma ou mais famílias grandes que incluam vários indivíduos afetados por determinada doença ao longo de várias gerações. A abordagem mais comum é basicamente um exercício de filogenética (o estudo de parentesco evolutivo entre grupos de organismos) e está profundamente enraizada em conceitos derivados da biologia evolutiva. É possível ilustrar descrevendo a busca de um gene envolvido na doença de Alzheimer. Cerca de 10% de todos os casos dessa doença nos Estados Unidos resultam de uma predisposição hereditária. Vários genes diferentes que foram descobertos, quando mutados, podem levar ao aparecimento precoce da doença de Alzheimer. Um desses genes (*PS1*) codifica a proteína presenilina-1 e para sua descoberta a análise de ligação foi muito utilizada.

Assim como os haplótipos dependem dos SNP que se encontram juntos em um cromossomo, a análise de ligação envolve a busca dos SNP que se encontram próximos do gene de interesse. Nesse tipo de estudo, os pesquisadores se concentram nas famílias afetadas pela doença, e procuram famílias nas quais as amostras de DNA podem ser obtidas a partir de indivíduos de várias gerações. As amostras de DNA são coletadas tanto dos membros da família afetados quanto dos não afetados. Em primeiro lugar, os pesquisadores localizam a região associada à doença em um cromossomo específico, utilizando conjuntos (chamados painéis) de localizações genômicas onde SNP comuns ou outras alterações genômicas mapeadas ocorrem em uma proporção significativa da população humana. Assim, muitos, mas não todos os humanos diferem na sequência genômica nesses locais. Utilizando um painel que inclua vários *loci* SNP bem caracterizados mapeados para cada cromossomo, os investigadores comparam os genótipos de indivíduos com e sem a doença, se concentrando principalmente nos familiares próximos. Para a doença de Alzheimer, duas das muitas linhagens familiares utilizadas para procurar esse gene no início da década de 1990 são mostradas na **Figura 9-34**. Utilizando essas linhagens, os investigadores descobriram variantes SNP específicas herdadas no mesmo ou quase no mesmo padrão que o gene causador da doença. O gene responsável pode gradualmente ser localizado em um único cromossomo, se a herança das variantes específicas de SNP no cromossomo se refletir perto da herança da condição da doença. Uma localização mais detalhada de um gene em um cromossomo causador de doença se baseia em métodos estatísticos para correlacionar a herança de polimorfismos adicionais mais próximos no espaço com a ocorrência da doença, se concentrando sobre um painel mais denso de polimorfismos que se sabe que ocorre no cromossomo de interesse. Quanto mais próximo um marcador está localizado do gene de uma doença, maior a probabilidade dele ter sido herdado junto com esse gene. Esse processo pode indicar uma

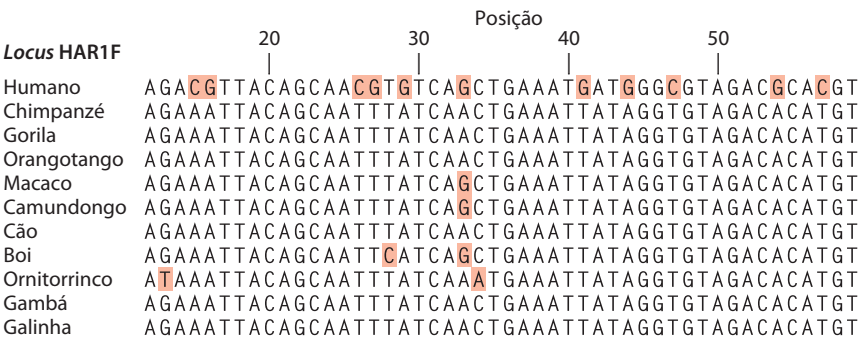


FIGURA 9-33 Evolução acelerada em alguns genes humanos. O *locus* HAR1F especifica um RNA não codificador que é altamente conservado em vertebrados. Em humanos, o gene HAR1F exibe um número incomum de mutações (vermelho-claro sombreado), fornecendo evidência de evolução acelerada.

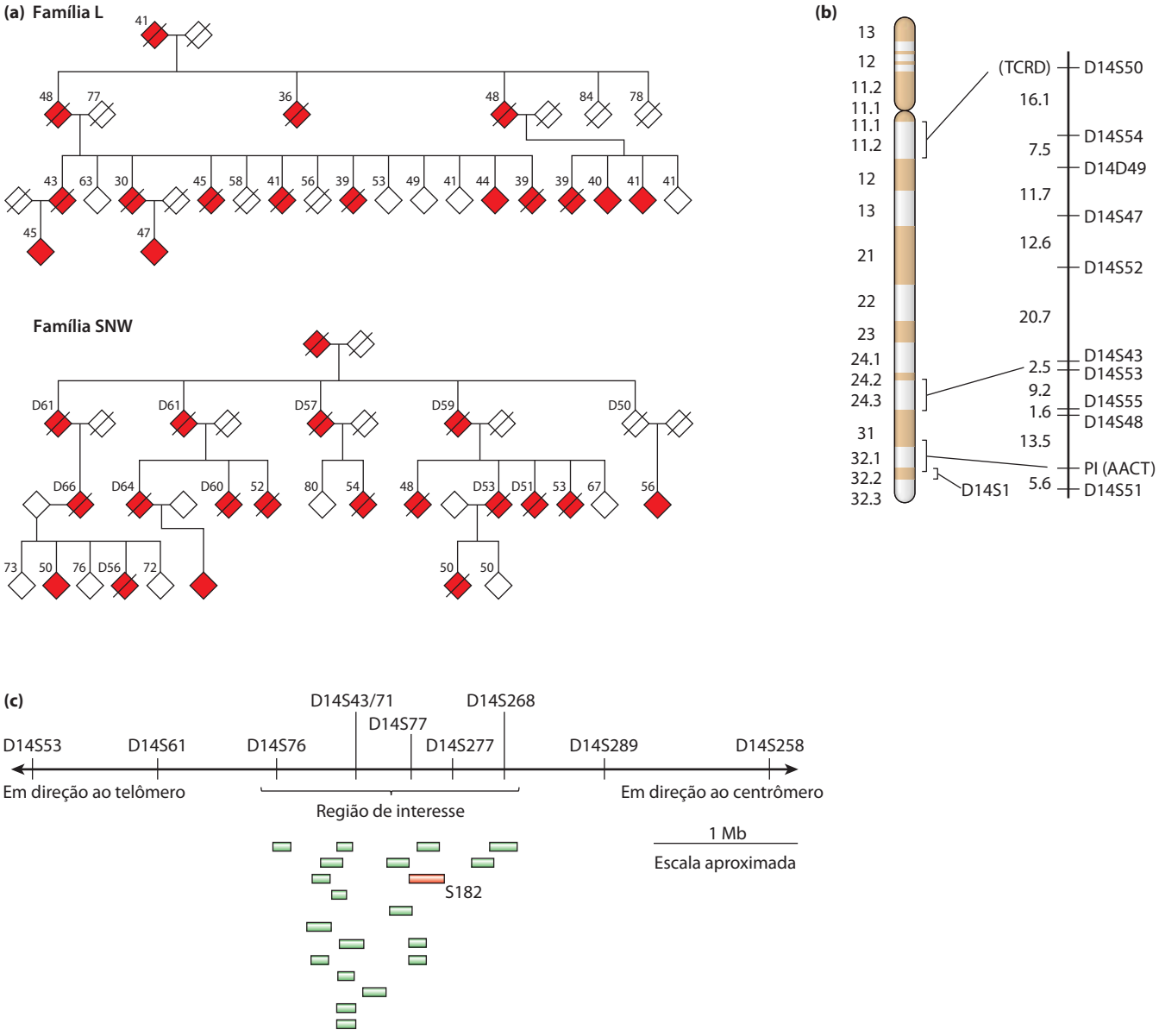


FIGURA 9-34 Análise de ligação na descoberta de genes de doenças. (a) Estes heredogramas para duas famílias afetadas pelo início precoce da doença de Alzheimer se baseiam nos dados disponíveis no momento do estudo. Os símbolos vermelhos representam os indivíduos afetados; as barras indicam as mortes antes ou logo no início do estudo. O número acima de cada símbolo corresponde à idade da pessoa, tanto no tempo do estudo quanto no momento da morte (indicada com a letra D). Para proteger a privacidade da família, o gênero não é indicado. **(b)** Cromossomo 14, com bandas criadas por alguns corantes. As posições dos marcadores de cromossomos são mostradas à direita, com a distância genética entre eles apresentada em uma medida de distância genética chamada de centimorgans, refletindo a frequência de recombinação entre eles. O *TCRD* (receptor delta de célula T) e *PI* (AACT [α1-antiquimotripsina]) são genes com alterações na população humana que foram utilizados como marcadores, juntamente com os SNP, no mapeamento de cromossomos. **(c)** A partir da comparação do DNA dos membros de famílias afetadas e não afetadas, uma região de interesse que contém 19 genes expressos foi finalmente definida próxima do marcador D14S43. O gene marcado *S182* (em vermelho) codifica a presenilina-1.

região do cromossomo que contém o gene. No exemplo da doença de Alzheimer, a análise de ligação indicou que o gene causador da doença estava em algum lugar próximo a um *locus* de SNP chamado D14S43 (Figura 9-34c).

As etapas finais da pesquisa para o gene da doença utilizam novamente o banco de dados do genoma humano. A região local que contém o gene é examinada e os seus genes são identificados. O DNA de muitos indivíduos, alguns com a doença e outros sem a doença, é sequenciado nessa região. Com o número crescente de indivíduos analisados, esse processo gradual leva à identificação das variantes genéticas consistentemente presentes em indivíduos com a doença e ausentes em indivíduos não afetados. A pesquisa pode ser auxiliada por uma compreensão da função dos genes na região-alvo, porque vias metabólicas específicas podem ser mais propensas a produzir a doença do que outras. Em 1995, o gene no cromossomo 14 associado à doença de Alzheimer foi identificado como o gene *S182*. O produto desse gene foi denominado de presenilina-1, e o próprio gene depois foi renomeado para *PS1*.

Mais complexos são os casos em que a doença é causada pela presença de mutações em dois genes diferentes (nenhum dos quais, isoladamente, é capaz de causar a doença), ou quando uma condição específica é potencializada por uma mutação, de outra maneira inócua, em outro gene. A identificação dos genes e mutações responsáveis por essas doenças digênicas é extremamente difícil e essas doenças, algumas vezes, só são possíveis de serem documentadas dentro de populações pequenas, isoladas e altamente puras.

Os bancos de dados do genoma abrem caminhos alternativos para a identificação de genes de doenças, especialmente quando a informação bioquímica sobre a doença é conhecida. No caso da doença de Alzheimer, um acúmulo de proteína β amiloide no sistema límbico e nos córtices de associação do cérebro é, pelo menos parcialmente, responsável pelos sintomas. Defeitos na presenilina-1 (e em uma proteína relacionada, a presenilina-2, codificada por um gene no cromossomo 1) conduzem a níveis corticais elevados de proteína β amiloide. Bancos de dados se concentram em catalogar tal informação funcional sobre os produtos proteicos de genes e nas redes de interação de proteínas, localização de SNP e outros dados, proporcionando um caminho rápido para a identificação de genes candidatos a determinada doença. Um pesquisador com certo conhecimento sobre os tipos de enzimas ou outras proteínas que contribuem para a doença pode utilizar esses bancos de dados para gerar uma lista de genes conhecidos que codificam proteínas com funções relevantes, genes adicionais não caracterizados com relações de ortólogos ou parálogos aos genes nesta lista, uma lista de proteínas conhecidas que interagem com as proteínas-alvo ou ortólogos em outros organismos e um mapa de posições de genes. Com os dados de linhagens de famílias selecionadas, uma pequena lista de genes potencialmente relevantes pode muitas vezes ser rapidamente determinada.

Essas abordagens não se limitam a doenças humanas. Os mesmos métodos podem ser utilizados para identificar os genes envolvidos em doenças – ou genes que produzem características desejáveis – em outros animais e plantas. ■

Sequências no genoma informam sobre o passado humano e fornecem oportunidades para o futuro

Cerca de 70.000 anos atrás, um pequeno grupo de seres humanos na África atravessou o Mar Vermelho em direção à Ásia. Talvez incentivados por alguma inovação na construção de pequenos barcos, ou conduzidos por conflitos ou fome, ou simplesmente curiosos, eles atravessaram a barreira de água. Essa colonização inicial, talvez envolvendo 1.000 indivíduos, começou uma jornada que só parou quando os seres humanos chegaram à Terra do Fogo (no extremo sul da América do Sul) muitos milhares de anos mais tarde. No processo, uma população estabelecida a partir de uma expansão de hominídeos na Eurásia, incluindo o *Homo neanderthalensis*, foi deslocada. Os neandertais desapareceram assim como o *Homo erectus* e outras linhas de hominídeos antes deles.

A história do aparecimento dos seres humanos modernos pela primeira vez na África algumas centenas de milhares de anos atrás, e suas migrações à medida que eles por fim se irradiaram para fora de África, está escrita no DNA. Com a utilização de sequências genômicas de várias espécies, a evolução de primatas e hominídeos se tornou mais nítida. Utilizando haplótipos presentes em populações humanas existentes, é possível traçar as migrações dos intrépidos antepassados humanos em todo o planeta (**Figura 9-35**).

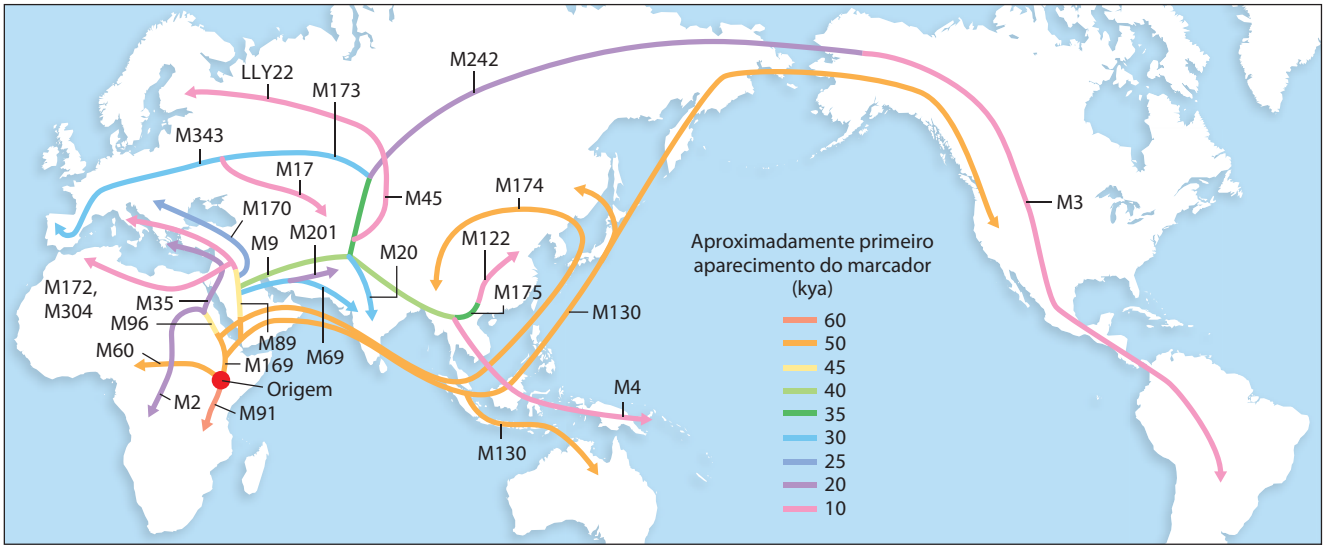


FIGURA 9-35 Os caminhos das migrações humanas. Quando uma pequena parte de uma população humana migra para fora de um grande grupo, ela carrega apenas parte de toda sua diversidade genética. Assim, alguns haplótipos estarão presentes no grupo que migrou, mas não todos eles. Ao mesmo tempo, mutações podem criar novos haplótipos ao longo do tempo. Este mapa foi gerado a partir de análises de marcadores genéticos (haplótipos definidos pelos números M ou LLY) no cromossomo Y. As amostras genéticas foram colhidas de populações indígenas há muito estabelecidas em pontos geográficos ao longo das rotas assinaladas. Os haplótipos que aparecem subitamente ao longo do caminho de migração, refletindo novas mudanças (mutações) em localizações genômicas de SNP específicas em algumas populações isoladas, são chamados de "eventos fundadores". Eles permitem aos pesquisadores rastrearem as migrações a partir desse ponto, uma vez que outras populações com o novo haplótipo foram provavelmente descendentes da população fundadora. A abreviação *kya* significa "milhares de anos atrás."

QUADRO 9-3 MÉTODOS Conhecendo os neandertais

Os seres humanos modernos e os neandertais conviveram na Europa e Ásia há relativamente recentes 30 mil anos. As populações ancestrais humanas e neandertais divergiram há cerca de 370.000 anos, antes do aparecimento dos seres humanos anatomicamente modernos. Os neandertais utilizavam ferramentas, viviam em pequenos grupos e enterravam seus mortos. Dos parentes hominídeos conhecidos dos humanos modernos, os neandertais são os mais próximos. Durante centenas de milênios, eles habitaram grande parte da Europa e Ásia Ocidental (Figura Q-1). Se o genoma do chimpanzé informe algo sobre o que é ser humano, talvez o genoma neandertal informe mais. Ossos enterrados que permanecem sendo retirados de cemitérios são fragmentos de DNA do genoma neandertal. As tecnologias desenvolvidas para uso em ciência forense (ver Quadro 9-1) e estudos de DNA arqueológico foram combinados para iniciar um projeto genoma neandertal.

Esse esforço é diferente dos projetos genoma que visam espécies existentes. O DNA neandertal está presente em pequenas quantidades contaminadas com o DNA de outros animais e de bactérias. Como se pode chegar até ele e como estar certo de que as sequências são realmente de neandertais? As respostas foram reveladas por aplicações inovadoras da biotecnologia. Essencialmente, as pequenas quantidades de fragmentos de DNA encontradas em um osso neandertal ou outros restos são clonadas em uma biblioteca, e os segmentos de DNA são sequenciados de forma aleatória, inclusive os contaminantes. Os resultados do sequenciamento são comparados com o banco de dados do genoma humano e do chimpanzé existentes. Os segmentos derivados de DNA neandertal são rapidamente diferenciados dos segmentos derivados de bactérias ou insetos por análise computadorizada, porque eles têm sequências estreitamente relacionadas ao DNA humano e do chimpanzé. Quando uma coleção de segmentos de DNA neandertal

FIGURA Q-1 Os neandertais ocuparam grande parte da Europa e Ásia Ocidental até cerca de 30.000 anos atrás. Os principais sítios arqueológicos neandertais estão aqui mostrados. (Observe que o grupo foi assim chamado por causa do sítio em Neanderthal na Alemanha.)

Os neandertais não foram simplesmente deslocados. Ocorreu alguma mistura. Utilizando-se métodos sensíveis com base em PCR, agora está disponível uma sequência quase completa do genoma de neandertais (Quadro 9-3). Sabe-se que cerca de 4% dos genomas humanos de não africanos são derivados de neandertais. Algumas populações humanas também adquiriram o DNA genômico de outro grupo recentemente descoberto, os denisovanos. O DNA de neandertais forneceu aos humanos um sistema imune mais complexo, tornando-nos mais resistentes à infecção, mas também um pouco mais suscetíveis a doenças autoimunes.

A história do passado humano está gradualmente tomando forma à medida que mais genomas humanos, dos vivos hoje e daqueles que viveram em milênios passados, estão sendo reunidos.

A promessa médica de sequências genômicas pessoais cresce à medida que mais genes subjacentes a doenças hereditárias são definidos. O conhecimento de sequências genômicas também oferece a possibilidade de alterá-los. Atualmente é comum modificar sequências de DNA de organismos a partir de bactérias e leveduras para plantas e mamíferos para fins de investigação e comercialização. Os esforços para

é sequenciada, pode ser utilizada como sonda para identificar os fragmentos de sequência em amostras antigas que se sobrepõem a esses fragmentos conhecidos. O potencial problema de contaminação com o DNA humano moderno estreitamente relacionado pode ser controlado pela análise do DNA mitocondrial. As populações humanas têm haplótipos facilmente identificáveis (conjuntos distintos de diferenças genômicas; ver Figura 9-30) no seu DNA mitocondrial, e a análise de amostras de neandertais mostrou que o DNA mitocondrial neandertal tem seus próprios haplótipos distintos. A presença nas amostras neandertais de algumas diferenças encontradas em pares de bases no banco de dados do chimpanzé, mas não no banco de dados humano, é mais uma evidência de que as sequências de hominídeos não humanos foram encontradas.

A conclusão dessa tarefa desafiadora está no horizonte. O projeto de sequenciamento do genoma neandertal, revelado no início de 2009, abrangeu mais de 60% das sequências genômicas. O sequenciamento final vai exigir um pouco mais de tempo. Os dados fornecem evidências de que os humanos modernos e os neandertais que foram a fonte desse DNA compartilharam um ancestral comum cerca de 700.000 anos atrás (Figura Q-2). A análise do DNA mitocondrial sugere que os dois grupos continuaram no mesmo caminho, com algum fluxo gênico entre eles, ao longo de mais 300 mil anos. As linhagens se separaram com a aparência de humanos anatomicamente modernos, embora hoje existam evidências para algum entrelaçamento das linhagens um pouco mais tarde.

Bibliotecas expandidas de DNA neandertal de diferentes conjuntos de restos mortais devem finalmente permitir uma análise da diversidade genética neandertal e, talvez, migrações de neandertais, oferecendo um fascinante olhar sobre o nosso passado hominídeo.

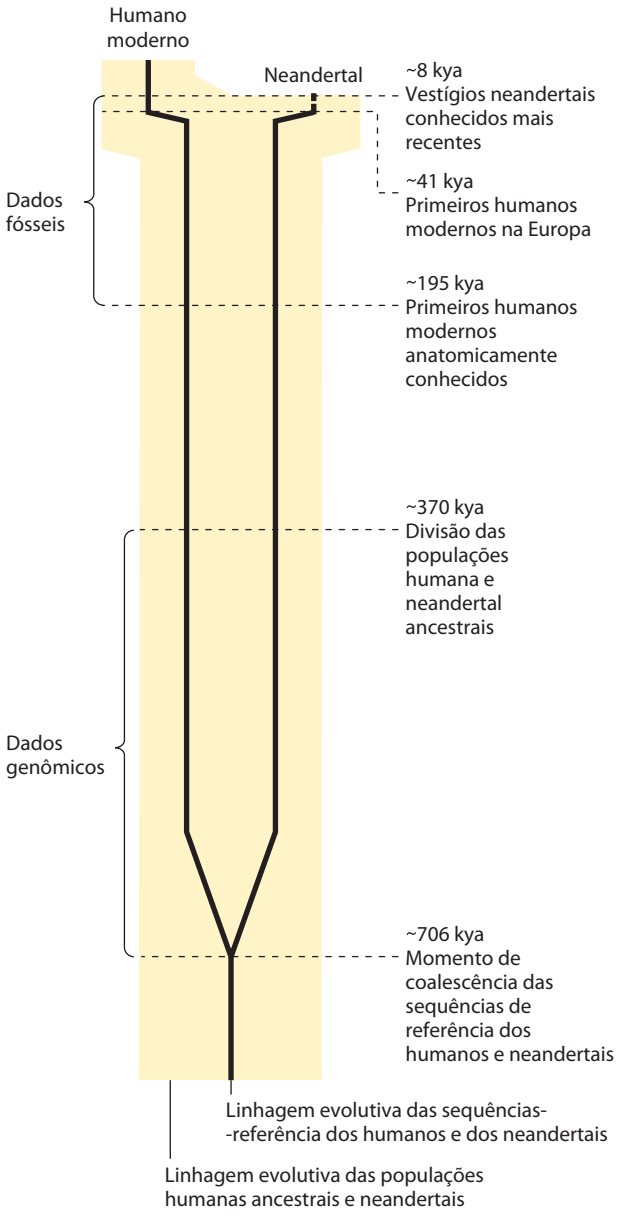


FIGURA Q-2 Esta linha do tempo mostra a divergência das sequências genômicas (linhas pretas) humanas e neandertais e das populações humanas ancestrais e neandertais (em amarelo). Os dados genômicos fornecem evidências para alguma intermistura das populações até cerca de 45.000 anos atrás. Eventos-chave na evolução humana são marcados.

curar doenças humanas hereditárias com terapia genética ainda não atingiram todo seu potencial, mas tecnologias para o fornecimento de genes estão constantemente se aperfeiçoando. Poucas disciplinas científicas afetarão o futuro da espécie humana mais do que a genômica moderna.

RESUMO 9.3 Genômica e história da humanidade

- Métodos de sequenciamento de última geração têm reduzido muito o tempo necessário para gerar sequências genômicas completas.
- Cerca de 30% do DNA no genoma humano estão nos éxons e íntrons dos genes que codificam proteínas. Cerca de metade do DNA é derivada de transposons parasitas. Muito do restante codifica RNA de muitos tipos. Sequências repetidas simples compõem o centrômero e telômeros.
- As alterações de genes que definem a humanidade podem ser discernidas em parte pela genômica comparativa utilizando-se outros primatas.
- A genômica comparativa também é utilizada para localizar as alterações no gene que definem doenças here-

ditárias e pode ser utilizada para estudar a evolução e a migração de nossos ancestrais humanos ao longo de milênios.

Termos-chave

Termos em negrito são definidos no glossário.

genômica 313
biologia de sistemas 313
clonagem 314
vetor 314
DNA recombinante 314
engenharia genética 314
endonucleases de restrição 314
DNA-ligases 314
plasmídeo 317
cromossomo artificial bacteriano (BAC) 319
cromossomo artificial de leveduras (YAC) 319
vetor de expressão 321
baculovírus 323
bacmídeo 323
mutagênese sítio-direcionada 323
proteína de fusão 325
marcador 325
reação em cadeia da polimerase (PCR) 327
repetições curtas em tandem (STR) 329
PCR quantitativa (qPCR) 331
biblioteca de DNA 332
biblioteca genômica 332
DNA complementar (cDNA) 332
biblioteca de cDNA 332
genômica comparativa 333
ortólogos 333
parálogos 333
sintenia 333
marcador de epítopo 333
análise de duplo-híbrido em leveduras 336
microarranjo de DNA 337
contig 342
polimorfismo de nucleotídeo único (SNP) 344
haplótipo 345

Leituras adicionais

Geral

Jackson, D.A., Symons, R.H., & Berg, P. (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose óperon of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 69, 2904-2909.

O primeiro experimento com DNA recombinante ligando o DNA de duas espécies.

Lobban, P.E. & Kaiser, A.D. (1973) Enzymatic end-to-end joining of DNA molecules. *J. Mol. Biol.* 78, 453-471.

Relato do primeiro experimento com DNA recombinante.

Estudo dos genes e seus produtos

Arnheim, N. & Erlich, H. (1992) Polymerase chain reaction strategy. *Annu. Rev. Biochem.* 61, 131-156.

Foster, E.A., Jobling, M.A., Taylor, P.G., Donnelly, P., de Knijff, P., Mieremet, R., Zerjal, T., & Tyler-Smith, C. (1999) The Thomas Jefferson paternity case. *Nature* 397, 32.

Último artigo de uma série de um interessante estudo de caso sobre o uso da biotecnologia para lidar com questões históricas.

Giepmans, B.N.G., Adams, S.R., Ellisman, M.H., & Tsien, R.Y. (2006) The fluorescent toolbox for assessing protein location and function. *Science* 312, 217-224.

Kayser, M. & de Knijff, P. (2011) Improving human forensics through advances in genetics, genomics, and molecular biology. *Nat. Rev. Genet.* **12**, 179-192.

Ståhl, P.L. & Lundeber, J. (2012) Toward the singlehour high-quality genome. *Annu. Rev. Biochem.* **81**, 359-378.

Zhao, J. & Grant, S.F.A. (2011) Advances in whole genome sequencing technology. *Curr. Pharm. Biotechnol.* **12**, 293-305.

Utilização de métodos com base no DNA para a compreensão das funções das proteínas

Budowle, B., Johnson, M.D., Fraser, C.M., Leighton, T.J., Murch, R.S., & Chakraborty, R. (2005) Genetic analysis and attribution of microbial forensics evidence. *Crit. Rev. Microbiol.* 31, 233-254.

Como a biotecnologia é utilizada contra o bioterrorismo.

Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**, 309-338.

Boa descrição de parte da genômica comparativa básica.

Stoughton, R.B. (2005) Applications of DNA microarrays in biology. *Annu. Rev. Biochem.* **74**, 53-82.

Terpe, K. (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *App. Microbiol. Biotechnol.* **72**, 211-222.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., et al. (2007) The *Sorcerer II* global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16.

Artigo que compõe uma série originada de um esforço ambicioso para provar a biodiversidade microbiana em todos os oceanos do mundo.

Genômica e história da humanidade

Alkan, C., Sajjadian, S., & Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods.* 8, 61-65.

Callaway, E. (2011) Ancient DNA reveals secrets of human history. *Nature* **476**, 136-137.

Carr, P.A. & Church, G.M. (2009) Genome engineering. *Nat. Biotechnol.* **27**, 1151-1162.

Carroll, S.B. (2003) Genetics and the making of *Homo sapiens*. *Nature* **422**, 849-857.

Gonzaga-Jauregui, C., Lupski, J.R., & Gibbs, R.A. (2012) Human genome sequencing in health and disease. *Annu. Rev.Med.* **63**, 35-61.

Green, E.D. & Guyer, M.S. (2011) Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204-213.

Kay, M.A. (2011) State-of -the-art gene-based therapies: the road ahead. *Nat. Rev. Genet.* **12**, 316-328.

Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature* **470**, 187-197.

Metzker, M.L. (2010) Sequencing technologies: the next generation. *Nat. Rev. Genet.* **11**, 31-46.

Perkel, J.M. (2011) Synthetic genomics: building a better bacterium. *Science* **331**, 1628-1630.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L., et al. (2010) Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature* 468, 1053-1060.

Os neandertais não foram os únicos outros grupos de hominídeos que os humanos encontraram em sua viagem através do continente asiático.

Stoneking, M. & Krause, J. (2011) Learning about human population history from ancient and modern genomes. *Nat. Rev. Genet.* **12**, 603-614.