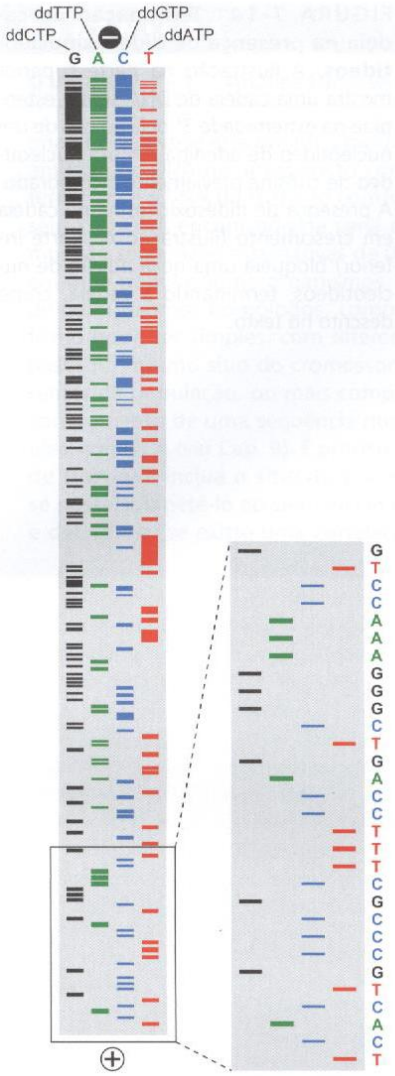


FIGURA 7-16 Gel de sequenciamento de DNA. Os comprimentos das cadeias de DNA, terminadas pelos dideoxinucleotídeos indicados na parte superior de cada linha, são determinados pela migração em gel de poliacrilamida, como mostrado. A leitura do gel, de baixo para cima, revela a sequência de 5′ para 3′.

gera um padrão semelhante a uma escada de bandas, cada degrau representando um C na fita-molde (Fig. 7-15b). Se, de forma semelhante, ddCTP, ddATP e ddTTP forem adicionados às reações de síntese de DNA, um a cada subconjunto de reação, então, no total, serão gerados quatro conjuntos de fragmentos de término, que, juntos, fornecerão a sequência completa de nucleotídeos do DNA. Para permitir a leitura da sequência, os fragmentos gerados em cada uma das quatro reações são resolvidos em gel de poliacrilamida (Fig. 7-16).

Como será visto a seguir, essa abordagem conceitualmente simples, desenvolvida inicialmente para sequenciar pequenos fragmentos de DNA definidos, sofreu uma série de adaptações e aperfeiçoamentos técnicos, que permitiram a análise de genomas inteiros (ver Quadro 7-2, Os sequenciadores são utilizados para o sequenciamento em larga escala).

Sequenciamento de um genoma bacteriano pelo método de *shotgun*

A bactéria *Haemophilus influenzae* foi o primeiro organismo de vida livre a ter sua sequência genômica determinada de forma completa e anotada. Foi uma escolha lógica, uma vez que essa bactéria tem um genoma compacto e pequeno, formado por apenas 1,8 milhão de pares de bases (Mb) de DNA (menos de 1:1.000º do tamanho do genoma humano). O genoma de *H. influenzae* foi digerido em vários fragmentos aleatórios com tamanho médio de 1 kb. Essas porções de DNA genômico foram clonados em um vetor de DNA plasmidial para criar uma biblioteca. O DNA foi preparado a partir de colônias de DNA recombinantes individuais e sequenciado separadamente em Sequenciadores, utilizando-se o método dos terminadores, discutido anteriormente neste capítulo. Esse método é chamado de sequenciamento por "*shotgun*". Colônias aleatórias de DNA recombinante são coletadas, processadas e sequenciadas. Para assegurar que cada nucleotídeo do genoma foi incluído na organização final do genoma, algo em torno de 30 mil a 40 mil clones recombinantes separados foram sequenciados. No total, cerca de 20 Mb de sequência genômica bruta foram sequenciadas (600 pb de sequência são obtidos em média por reação, e 600 pb × 33.000 colônias diferentes = 20 Mb de sequência de DNA total). Essa é chamada de uma **cobertura de sequência de 10 ×**. A princípio, isso significa que cada nucleotídeo no genoma foi sequenciado 10 vezes.

Esse método pode parecer enfadonho, mas é consideravelmente mais rápido e menos oneroso do que as técnicas originalmente desenvolvidas. Uma estratégia inicial necessitava do sequenciamento sistemático de cada fragmento de restrição de DNA definido em um mapa físico do cromossomo bacteriano. Uma desvantagem desse procedimento é que a maioria dos fragmentos de restrição conhecidos era maior do que a quantidade de informação de sequência de DNA que podia ser gerada por uma única reação. Consequentemente, seriam necessários ciclos adicionais de digestão, mapeamento e sequenciamento para que fosse obtida uma sequência completa de qualquer região determinada do genoma. Estas etapas adicionais de clonagem e mapeamento de restrição são consideravelmente mais demoradas do que o sequenciamento automatizado repetitivo de fragmentos aleatórios de DNA. Em outras palavras, o computador é muito mais rápido na organização das sequências aleatórias do que o tempo necessário para realizar um mapeamento de restrição em pequena escala do cromossomo bacteriano.

As cerca de 30.000 leituras de sequenciamento derivadas de fragmentos aleatórios de DNA genômico são diretamente inseridas no computador, e são utilizados programas para unir sequências de DNA sobrepostas. Esse processo é conceitualmente semelhante à montagem de um gigantesco jogo de palavras cruzadas denso no qual as palavras determinadas dão pistas acerca das palavras sobrepostas, porém, desconhecidas. Os fragmentos aleatórios de DNA são "encaixados" de acordo com as sequências que se emparelham. A montagem sequencial dessas pequenas sequências de DNA finalmente resulta em uma montagem contínua única, também chamada de contig (ver Fig. 7-18).

EXPERIMENTOS-CHAVE

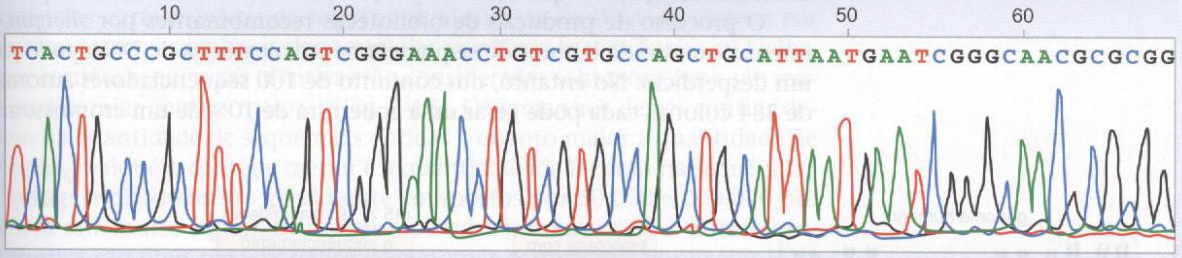
Quadro 7-2 Os sequenciadores são utilizados para o sequenciamento em larga escala

Quando o sequenciamento do genoma humano foi primeiramente imaginado, parecia ser uma iniciativa desanimadora, praticamente sem chance de êxito. Afinal, o genoma humano completo consiste em impressionantes 3 bilhões de pares de bases (3×10^9), e os primeiros métodos para a determinação da sequência de nucleotídeos, mesmo de pequenos fragmentos de DNA, eram lentos. Na década de 1980 e no início da de 1990, um único pesquisador poderia produzir apenas algumas centenas de pares de bases, talvez 500 pb, de sequência de DNA em um dia ou dois de esforço concentrado. Várias inovações tecnológicas aceleraram de forma significativa a velocidade e a confiabilidade do sequenciamento de DNA.

Como descrito na seção anterior, o método de terminação de cadeia produz subconjuntos de DNA que diferem em tamanho por apenas um nucleotídeo. Inicialmente, eram necessários longos géis de poliacrilamida para separar esses segmentos de DNA (ver Fig. 7-16). Entretanto, nos últimos anos, esses géis inconvenientes e de difícil manuseio foram substituídos por pequenas colunas, que permitem a resolução de DNA em apenas 2 a 3 horas. Essas pequenas colunas reutilizáveis permitem o fracionamento de fragmentos de DNA de 700 a 800 pb, similar à capacidade dos géis de poliacrilamida mais longos que elas substituíram.

O principal avanço técnico no sequenciamento de DNA foi resultado do uso de **nucleotídeos terminadores de cadeia fluorescentes**. A princípio, é possível marcar cada um dos segmentos de DNA de um fragmento com uma única "cor". A cor de cada fragmento de DNA depende da identificação do último nucleotídeo. Por exemplo, os DNAs que terminam com um resíduo de T na posição 50 (e em todas posições terminadas pelo T didesoxi) do DNA-molde são marcados em vermelho, enquanto os DNAs que terminam com um resíduo de G na posição 51 (e em todas posições terminadas pelo G didesoxi) poderão ser marcados em preto. Assim, cada segmento de DNA tem uma única cor e tamanho. À medida que eles são fracionados nas colunas de sequenciamento de acordo com o tamanho, sensores fluorescentes detectam a cor de cada fragmento de DNA (Quadro 7-2, Fig. 1). Dessa maneira, uma única coluna produz de 600 a 800 pb de sequência de DNA em menos de 3 horas de separação por tamanho.

As máquinas de sequenciamento automatizadas – **Sequenciadores** – foram desenvolvidas com 384 diferentes colunas de fracionamento. A princípio, essas máquinas podem gerar mais de 200.000 nucleotídeos (200 kb) de sequências brutas de DNA em poucas horas. Em um dia de 9 horas, cada máquina pode produzir três "corridas" de sequenciamento e mais de meia megabase (500 kb) de informações de sequência. Um conjunto de 100 dessas máquinas poderia gerar o equivalente a um genoma humano, 3×10^9 pb, em apenas 2 meses. Atualmente, há cinco grandes centros de sequenciamento nos Estados Unidos e no Reino Unido. Cada um deles tem grandes conjuntos de máquinas de sequenciamento automatizadas. Juntos, esses cinco centros produzem o impressionante número de 60×10^9 pb de informação de sequências brutas de DNA por ano. Isso corresponde ao equivalente a 20 genomas humanos por ano! Mas como será visto mais tarde, isso é muito pouco quando comparado aos sequenciadores de última geração que produzem rotineiramente o equivalente a um genoma humano completo em uma única corrida de apenas algumas horas.



QUADRO 7-2 FIGURA 1 Leitura da sequência de DNA. Nesta reação, como descrito no texto, foram utilizados didesoxinucleotídeos marcados nas extremidades com fluorescência, e as cadeias foram separadas por cromatografia em coluna. O perfil das posições de As está representado em verde; de Ts, em vermelho; de Gs, em preto; e de Cs, em azul.

A estratégia de *shotgun* permite a montagem parcial de grandes sequências genômicas

Na discussão anterior, viu-se que o sequenciamento de fragmentos de DNA curtos, com até 600 pb, é incrivelmente rápido e eficiente. De fato, as máquinas de sequenciamento automatizadas são tão eficientes que ultrapassam em muito a capacidade de montar e anotar a informação de sequências brutas de DNA. Em outras palavras, a etapa limitante da velocidade na determinação da sequência completa de DNA de genomas complexos, como o genoma humano, é a análise dos dados e não a produção dos dados em si. Esse problema

está se tornando ainda mais grave à medida que os métodos de sequenciamento se tornam mais rápidos e mais poderosos. Hoje é possível gerar informações de sequências de vários bilhões de pares de bases (pares de gigabases, Gb) de DNA em uma "corrida" de uma máquina automatizada (ver seção intitulada O genoma humano de US$ 1.000 está ao nosso alcance). Agora, será considerado como o método de sequenciamento por *shotgun*, utilizado para determinar a sequência completa do genoma de *H. influenzae*, foi adaptado para genomas animais muito maiores e complexos.

Um cromossomo humano médio é composto por 150 Mb. Assim, a sequência de DNA de 600 pb fornecida por uma reação de sequenciamento normal representa apenas 0,0004% de um cromossomo típico. Consequentemente, para determinar a sequência completa do cromossomo, é necessário gerar um grande número de leituras de sequências a partir de vários fragmentos curtos de DNA (Fig. 7-17). Para que seja atingido esse objetivo, o DNA de cada um dos 23 cromossomos humanos que compõem o genoma humano é preparado e, então, distribuído em conjuntos ou bibliotecas de pequenos fragmentos, através de agulhas pressurizadas de pequeno calibre. A coleção de fragmentos pequenos, cada um deles derivado de cromossomos individuais, é então reduzida a conjuntos. Em geral, duas ou três bibliotecas são construídas para fragmentos de diferentes (crescentes) tamanhos – por exemplo, fragmentos de 1, 5 ou 100 kb de comprimento. Esses fragmentos são, então, aleatoriamente clonados em plasmídeos bacterianos, como já foi descrito.

O DNA recombinante, contendo uma porção aleatória de um cromossomo humano, pode ser rapidamente isolado de plasmídeos bacterianos e, então, rapidamente sequenciado, utilizando-se as máquinas de sequenciamento automatizadas. Para assegurar que todas as sequências do cromossomo sejam lidas, são processados, em média, 2 milhões de fragmentos aleatórios de DNA. Com uma média de 600 pb de sequência de DNA por fragmento, esse procedimento produz mais de 1 bilhão de pares de bases (1 Gb) de dados sequenciais, ou aproximadamente 10 vezes a quantidade média de DNA de um cromossomo normal. Como discutido anteriormente para o sequenciamento do cromossomo bacteriano, a leitura de 10 vezes a quantidade de sequência de um cromossomo garante que todas as porções do cromossomo serão capturadas.

O processo de produção de bibliotecas recombinantes por *shotgun* e um imenso excesso de leituras de sequenciamento aleatórias de DNA pode parecer um desperdício. No entanto, um conjunto de 100 sequenciadores automáticos de 384 colunas cada pode gerar uma cobertura de 10× de um cromossomo hu-

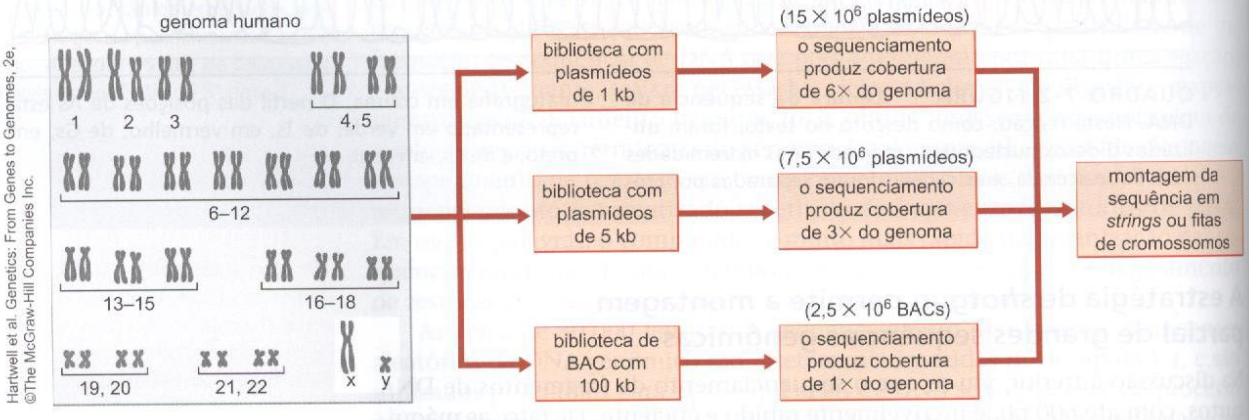


FIGURA 7-17 Estratégia para a construção e sequenciamento de bibliotecas de genomas inteiros. As sequências contíguas são determinadas pelo sequenciamento de *shotgun* de fragmentos curtos de DNA genômico. Os contigs são estendidos pela utilização das sequências terminais derivadas de fragmentos maiores nos insertos de 5 kb e 100 kb, como descrito no texto. BAC, cromossomo artificial de bactérias. (Adaptada, com permissão, de Hartwell L. et al. 2003. *Genetics: From genes to genomes*, 2nd ed., Fig. 10-13. © McGraw-Hill.)

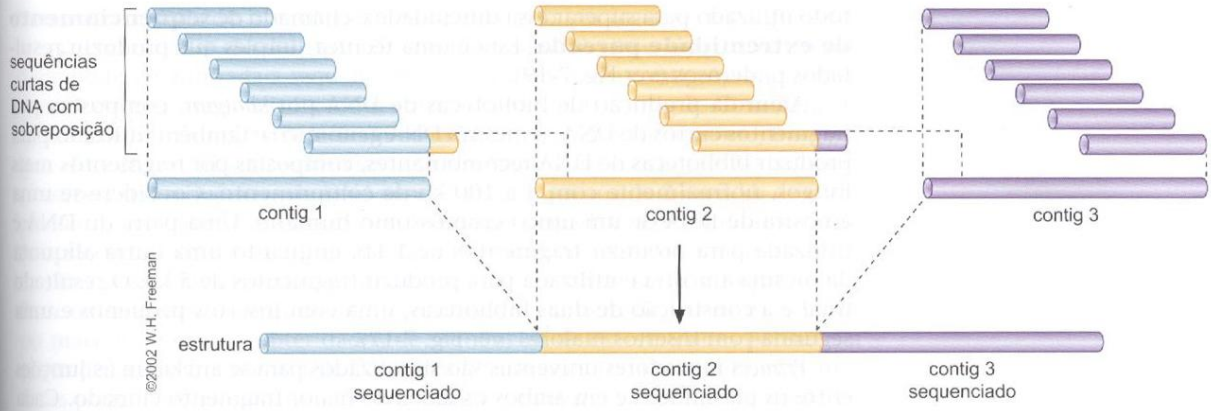


FIGURA 7-18 Os contigs são unidos pelo sequenciamento das extremidades de grandes fragmentos de DNA. Por exemplo, a extremidade de um fragmento aleatório de DNA genômico de 100 kb pode conter sequências equivalentes ao contig 1, enquanto a outra extremidade equivale a sequências no contig 2. Isso coloca os dois contigs sobre a mesma estrutura. (Adaptada, com permissão, de Griffiths A.J.F. et al. 2002. *Modern genetics*, 2nd ed., Fig. 9-29b. © W.H. Freeman.)

mano em poucas semanas. Isso é consideravelmente mais rápido do que os métodos que envolvem o isolamento de regiões conhecidas em um cromossomo e o sequenciamento de um conjunto conhecido de fragmentos de DNA que se situam lado a lado. Assim, o princípio tecnológico fundamental que facilitou o sequenciamento do genoma humano foi a confiabilidade do **sequenciamento automatizado por *shotgun*** e a subsequente utilização de computadores para a montagem dos diferentes fragmentos como um quebra-cabeça. A combinação de máquinas de sequenciamento automatizadas e computadores provou ser uma potente estratégia que permitiu a finalização do sequenciamento do genoma humano anos antes do prazo originalmente planejado.

Sofisticados programas de computador foram desenvolvidos para organizar as sequências curtas obtidas de sequências aleatórias de DNA de *shotgun* em sequências contínuas maiores, denominadas **contigs**. As sequências ou "leituras" que contêm sequências idênticas são consideradas sobreposições e são reunidas para formar contigs maiores (Fig. 7-18). Os tamanhos desses contigs dependem da quantidade de sequências obtidas – quanto maior a quantidade de sequências, maior é o contig e menor é a quantidade de lacunas na sequência.

Contigs individuais são geralmente compostos por 50.000 a 200.000 pb. Isso ainda está muito aquém de um cromossomo humano típico. Entretanto, esses contigs são úteis para a análise de genomas compactos. Por exemplo, o genoma de *Drosophila* contém, em média, um gene a cada 10 kb, de forma que um contig típico contém vários genes ligados. Infelizmente, os genomas mais complexos contêm, frequentemente, densidades gênicas muito menores (ver Cap. 8). O genoma humano contém, em média, um gene a cada 100 kb, de forma que um contig é normalmente insuficiente para acomodar um gene inteiro, e muito menos uma série de genes ligados. Será considerado, agora, como contigs relativamente curtos são montados, formando **estruturas** maiores, que apresentam de 1 a 2 Mb de comprimento.

A estratégia de extremidades pareadas permite a montagem de longas estruturas genômicas

Uma limitação importante na produção de contigs longos é a ocorrência de DNAs repetitivos (ver Cap. 8). Essas sequências dificultam o processo de montagem, uma vez que fragmentos aleatórios de DNA presentes em regiões não ligadas de um cromossomo ou genoma podem ser considerados uma sobreposição, devido à presença da mesma sequência de DNA repetitivo. Um mé-

todo utilizado para superar essa dificuldade é chamado de **sequenciamento de extremidade pareada**. Esta é uma técnica simples que produziu resultados poderosos (ver Fig. 7-19).

Além da produção de bibliotecas de DNA por *shotgun*, compostas por fragmentos curtos de DNA, o mesmo DNA genômico é também utilizado para produzir bibliotecas de DNA recombinantes, compostas por fragmentos mais longos, normalmente com 3 a 100 kb de comprimento. Considere-se uma amostra de DNA de um único cromossomo humano. Uma parte do DNA é utilizada para produzir fragmentos de 1 kb, enquanto uma outra alíquota da mesma amostra é utilizada para produzir fragmentos de 5 kb. O resultado final é a construção de duas bibliotecas, uma com insertos pequenos e uma segunda com insertos maiores (ver Fig. 7-17).

Primers iniciadores universais são sintetizados para se anelarem às junções entre os plasmídeos e em ambos os lados do maior fragmento clonado. Cada sequenciamento produzirá cerca de 600 pb de informação de sequência em

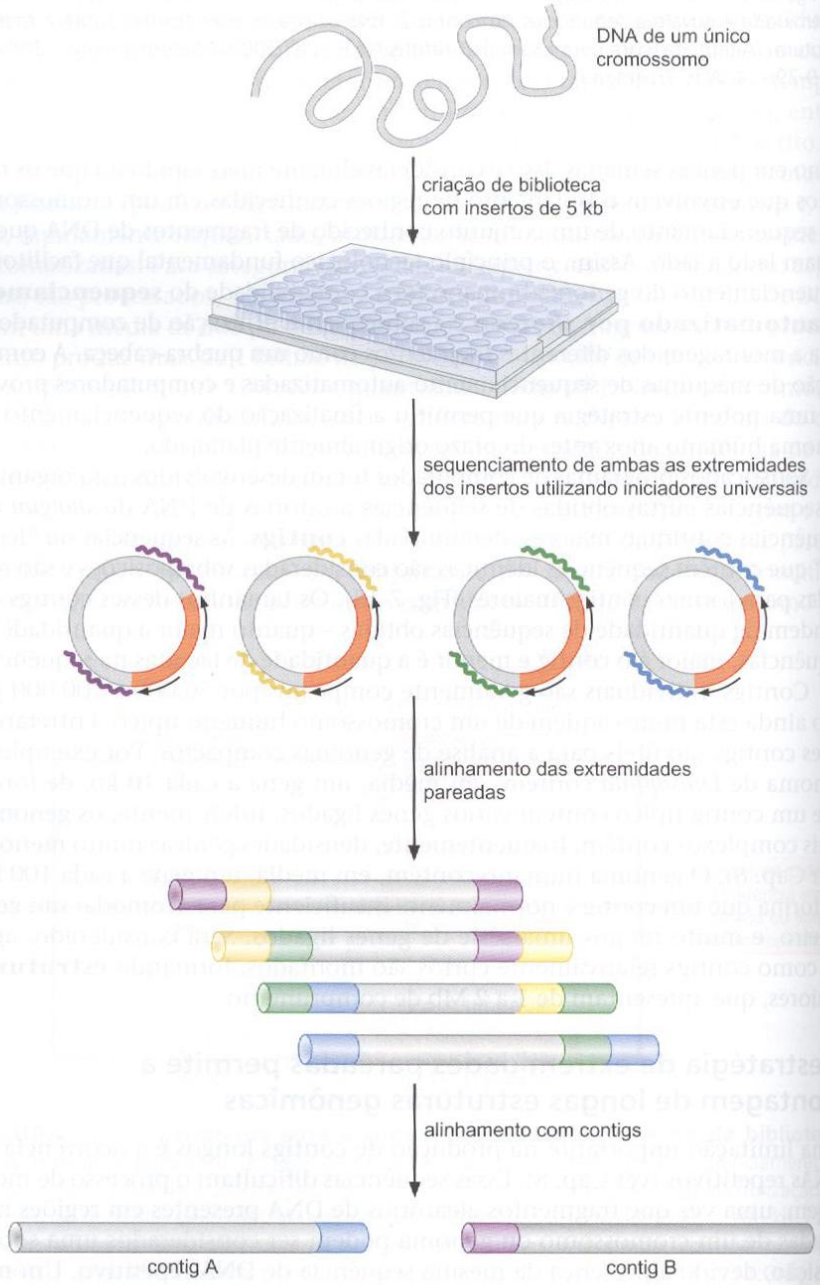


FIGURA 7-19 Biblioteca de "*shotgun*" contendo insertos aleatórios de DNA genômico de 5 kb de comprimento. Cada poço da placa contém um inserto diferente. Sequências de 600 pb de comprimento são determinadas para ambas as extremidades de cada DNA genômico (coloridos). Estas sequências de extremidades pareadas são utilizadas para alinhar diferentes contigs. Neste exemplo, o fragmento de DNA genômico de 5 kb com as sequências em azul contém sequências correspondentes ao contig A e ao contig B.

cada extremidade do inserto aleatório. É mantido um registro das sequências cujas extremidades são derivadas do mesmo fragmento inserido. Uma extremidade pode ser alinhada a sequências contidas dentro do contig A, enquanto a outra extremidade se alinha a um contig diferente, o contig B. Considera-se, agora, que os contigs A e B derivam da mesma região do cromossomo, uma vez que eles compartilham sequências com um fragmento comum de 5 kb. Como a maioria das sequências de DNA repetitivo tem menos de 2 ou 3 kb de comprimento, as sequências com as "extremidades pareadas" a partir do inserto de 5 kb são suficientes para cobrir contigs interrompidos por DNAs repetitivos.

Em geral, esses resultados produzem contigs < 500 kb de comprimento. Para obter dados de sequência mais longas, na ordem de várias megabases ou mais, é necessário obter dados de sequências de extremidades pareadas a partir de longos fragmentos de DNA com pelo menos 100 kb de comprimento. Estas podem ser obtidas usando um vetor de clonagem especial chamado de **BAC (cromossomo artificial de bactérias** [*bacterial artificial chromosome*]) que pode acomodar insertos muito grandes, de até centenas de quilobases de DNA. O princípio para produzir informação de sequência de longo alcance nesses vetores é o mesmo descrito para os insertos de 5 kb. Os iniciadores são utilizados para a obtenção de leituras de 600 pb a partir de ambas as extremidades do inserto clonado em BAC. Essas sequências são, então, alinhadas em diferentes contigs, os quais podem ser designados para a mesma estrutura, porque compartilham sequências de um inserto BAC comum. Frequentemente, a utilização de BACs permite a designação de múltiplos contigs em uma única estrutura, ou arcabouço, com várias megabases (ver Fig. 7-18).

O genoma humano de US$ 1.000 está ao nosso alcance

O sequenciamento dos dois primeiros genomas humanos (um do National Institutes of Health [NIH] e outro de uma companhia privada) custou mais de US$ 300 milhões. Hoje há uma campanha para utilizar a nanotecnologia para a produção de sequenciamento genômico rápido e barato. O objetivo é tornar a tecnologia rápida, simples e barata o suficiente para permitir o sequenciamento de genomas individuais para o diagnóstico clínico. A primeira geração de máquinas de sequenciamento por nanotecnologia de alto rendimento já está disponível.

O sequenciador 454 Life Sciences gera até 400 Mb de informações de sequência em uma "corrida" de 4 horas. O princípio básico é muito inteligente. Pequenos fragmentos de DNA (genômico, cDNA, etc.) são misturados com pequenas esferas. A mistura é suficientemente diluída para garantir que uma única molécula de DNA se ligue a uma única esfera. Em seguida, as esferas contendo DNA são dispersadas em uma placa de silicone que consiste em 400.000 poços com capacidade de picolitros, regularmente espaçados. O pequeno tamanho dos poços garante que cada um deles capture não mais do que uma única esfera. Realiza-se PCR diretamente nos DNAs acoplados às esferas para amplificar cada molécula de DNA (Fig. 7-20). Portanto, uma população homogênea de moléculas de DNA é criada em cada poço, sendo

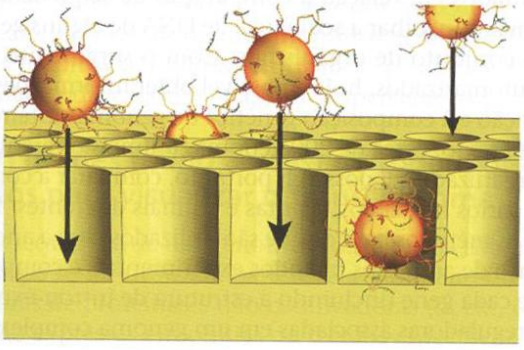


FIGURA 7-20 Desenho dos poros individuais no aparato de sequenciamento 454. Cada poro contém uma pequena esfera com uma sequência de DNA amplificada. Rodadas sequenciais de sequenciamento são detectadas pela liberação de pirofosfato e luz. Os detalhes desse método são descritos no texto. (Reproduzida, com permissão, de Margulies M. et al. 2005. *Nature* **437**: 376-380, Fig. 1a. © Macmillan.)

então utilizada como molde para uma rodada adicional de síntese de DNA. O sequenciamento é realizado por etapas com a placa sendo separadamente exposta a dATP, dGTP, dCTP e dTTP de maneira sequencial, com um ciclo de lavagem entre cada pulso de substrato de desoxinucleotídeo. A incorporação de um desoxinucleotídeo depende da presença da base complementar no molde e resulta na liberação de pirofosfato. Essa liberação promove uma reação enzimática que produz pulsos de luz, os quais são detectados por um microprocessador ligado a um computador. Os pulsos de luz indicam qual nucleotídeo é incorporado em cada poço durante cada rodada de síntese, produzindo, assim, a sequência de DNA contida em todos os 400.000 poços. A adição sequencial de cada nucleotídeo é continuada até que 200 a 250 bases tenham sido determinadas para cada fragmento de DNA.

O sequenciador 454 produziu o genoma completo do autor principal deste livro (por algum motivo, a companhia parece menos interessada nos genomas dos outros autores). A 100 Mb de sequência genômica por "corrida", a cobertura completa de 1× do genoma de Watson necessitou de apenas 30 corridas (2 a 3 semanas em uma máquina). Se iniciado agora (no momento da redação deste texto), o custo total estaria em torno de US$ 10.000 a US$ 30.000, uma pequena fração do custo da primeira sequência de genoma humano. A informação de sequência não é necessariamente suficiente para produzir uma montagem *de novo* do genoma. Em vez disso, a sequência finalizada do genoma humano produzida pelo NIH é utilizada como molde para comparações. Cada uma das leituras de sequência de 200 a 250 pb produzidas pelo sequenciador 454 são identificadas no genoma finalizado até que as variantes de cada gene de Watson sejam identificadas. Portanto, o significado do sequenciamento de um genoma humano mudou. Como tem-se uma montagem de sequência finalizada de genoma inteiro disponível, novos genomas necessitam apenas de curtas leituras de sequenciamento para obter um atlas abrangente da composição genética única de um indivíduo.

A nova geração de sequenciadores está se aproximando do objetivo do genoma de US$ 1.000. A Illumina produziu uma máquina que pode gerar centenas de milhões de leituras de sequências de 200 pb por corrida. O princípio básico é semelhante ao visto para o sequenciador 454 Life Sciences. A diferença é que moléculas de DNA individuais são ligadas a uma lâmina de vidro. Realiza-se uma amplificação limitada por PCR para produzir aproximadamente 1.000 cópias por molécula de DNA. Reações de síntese sequencial de DNA são realizadas e detectadas pela liberação de pirofosfato. Os sequenciadores Illumina produzem rotineiramente vários gigabases de informação de sequência de DNA em uma única corrida. Uma variedade de métodos de sequenciamento de alto rendimento de última geração está sendo desenvolvida, incluindo sequenciamento por semicondutor de íon, que detecta o íon hidrogênio liberado pela incorporação de um nucleotídeo durante a síntese de DNA.

GENÔMICA

Antes do advento do sequenciamento do genoma inteiro, os pesquisadores estavam muito limitados em relação à comparação de sequências de DNA. Eles conseguiam, no máximo, olhar a sequência de DNA de alguns genes individuais em um pequeno conjunto de organismos. Com o surgimento dos poderosos sequenciadores automatizados, hoje é possível obter informações completas em relação à organização e à composição genética de genomas inteiros. Na verdade, até a redação deste texto, cerca de 200 genomas animais diferentes haviam sido sequenciados e organizados. É possível, portanto, comparar a composição genética completa de vários micróbios, plantas e animais diferentes. Nesta seção, serão considerados os métodos básicos que são utilizados para a anotação de genomas – ou seja, o uso de ambos os métodos experimentais e computacionais para a identificação de cada gene (incluindo a estrutura de íntron-éxon; ver Cap. 14) e das sequências reguladoras associadas em um genoma complexo.

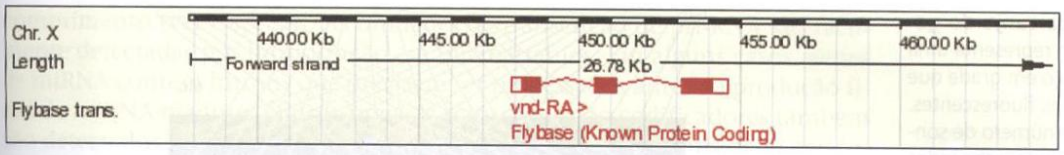


FIGURA 7-21 Estrutura do *locus vnd* em *Drosophila*. Intervalo de cerca de 25 kb no cromossomo X que contém o gene *vnd*. A unidade de transcrição de *vnd* contém três éxons e dois íntrons. As porções não preenchidas nos éxons 5′ (à esquerda) e 3′ (à direita) indicam sequências não codificadoras que não contribuem para o produto proteico final. FlyBase é um banco de dados padronizado, utilizado para analisar o genoma de *Drosophila*.

Ferramentas de bioinformática facilitam a identificação de genes codificadores de proteína no genoma inteiro

Montagens de sequências genômicas correspondem a blocos contíguos de milhões de As, Gs, Cs e Ts sequenciais que abrangem cada cromossomo do organismo em questão. Elas são grandes, tediosas e não informativas, a não ser que sejam "anotadas". Como descrito nas próximas páginas, **anotação** é a identificação sistemática de cada trecho de DNA genômico que contém informações de codificação de proteínas ou sequências não codificadoras que especificam RNAs reguladores, como os microRNAs (miRNAs; ver Cap. 20). A estrutura detalhada de íntron-éxon de cada unidade de transcrição é identificada, e nos casos em que o genoma em questão corresponde a um organismo-modelo (p. ex., levedura e mosca-da-fruta), é possível atribuir funções potenciais ou conhecidas à maioria dos genes do genoma. Apenas quando essa informação está disponível é possível catalogar a capacidade codificadora completa do genoma e comparar seu conteúdo aos de outros genomas.

Nos genomas de bactérias e de eucariotos simples, o processo de anotação de genes codificadores de proteínas é relativamente direto e corresponde, basicamente, à identificação de fases abertas de leitura (ORFs, *open reading frames*). Embora nem todas as ORFs – especialmente as pequenas – sejam genes codificadores de proteínas, este processo é bastante efetivo, e o desafio principal está em atribuir corretamente as funções dos genes.

Nos genomas animais, com estruturas complexas de íntron-éxon, o desafio é muito maior. Neste caso, diversas ferramentas de bioinformática são necessárias para a identificação de genes e a determinação da composição genética dos genomas complexos. Vários programas de computação foram desenvolvidos e são capazes de identificar genes codificadores potenciais, por meio de vários critérios de sequência (Fig. 7-21), incluindo a ocorrência de ORFs estendidas, flanqueadas por sítios de *splice* apropriados em 5′ e 3′. Conforme discutido no Capítulo 14, sítios de *splice* doadores e aceptores são sequências curtas e, de certa maneira, degeneradas, mas ainda assim ajudam a identificar as fronteiras entre éxon e íntron quando considerados no contexto de informações adicionais, como os dados de sequências de etiquetas de sequências expressas (EST, *expressed sequence tag*), que serão considerados mais adiante. Entretanto, esses métodos ainda não foram ajustados para atingir 100% de precisão. Talvez cerca de 75% de todos os genes possam ser identificados dessa maneira, mas ainda restam muitos sem identificação, e, mesmo entre os genes preditos que são identificados, pequenos éxons – particularmente éxons não codificadores – passam despercebidos.

Arranjos em grades (*tiling arrays*) de todo o genoma são utilizados para visualizar o transcriptoma

Uma vez que a sequência de um genoma inteiro tenha sido montada para um organismo, ela pode ser utilizada para revelar amplamente todas as sequências codificadoras de proteínas e não codificadoras (p. ex., íntrons e genes de miRNA) que são expressas em células ou tecidos específicos.

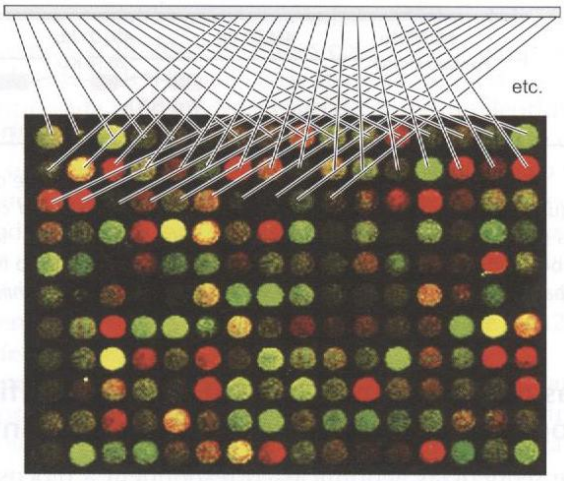


FIGURA 7-22 Microarranjo do genoma inteiro. A imagem representa uma porção de um microarranjo em grade que foi hibridizado com sondas fluorescentes. A grade inclui um grande número de sondas de DNA uniformemente espaçadas através de uma região de interesse (p. ex., um genoma inteiro).

A porção do genoma de um organismo que atua como molde para a síntese de RNA é conhecida como **transcriptoma**. Para identificar essa porção do genoma, DNAs de fita simples sintéticos com 50 nucleotídeos de comprimento são plotados em uma lâmina de vidro ou silicone. Geralmente, um oligonucleotídeo é produzido para cada 100 a 150 pb de sequência de DNA de maneira sequencial ao longo do genoma, resultando em um arranjo de sequências de DNA (microarranjos ou *tiling array*). A tecnologia de *tiling array* para o genoma inteiro está avançando rapidamente, e hoje é possível produzir arranjos completos em uma única lâmina de vidro ou *chip* de silicone com apenas 1 cm^2 de área. Por exemplo, 1 milhão de sequências de 50 nucleotídeos abrangem o genoma inteiro de *Drosophila*, e todos estes oligonucleotídeos podem ser plotados em um único *chip* de DNA. Cada ponto no *chip* (i.e., cada sequência de oligonucleotídeo) é tão pequeno que os sinais de hibridização são detectados por microssensores ligados a um microscópio, como será descrito mais adiante.

Para visualizar o transcriptoma, os microarranjos são hibridizados com sondas de RNA (ou cDNA) fluorescentemente marcadas (ver Fig. 7-22). Essas sondas podem ser derivadas de um tipo celular específico, como os músculos caudais da larva de ascídia ou células de levedura crescidas em um determinado meio de cultivo. O resultado final é uma série de sinais de hibridização sobrepostos em todas as sequências codificadoras de proteínas preditas ao longo do genoma (Fig. 7-23). Uma estratégia alternativa para o perfil de transcriptoma é o sequenciamento de alto rendimento de cDNA preparado a partir de células cultivadas ou de tecidos isolados.

Microarranjos de genoma inteiro fornecem informações imediatas em relação à estrutura de íntron-éxon de unidades de transcrição individuais (Fig. 7-23). Isso deve-se à natureza instável dos transcritos intrônicos. Embora o RNA total seja geralmente utilizado para estes experimentos, as sequências exônicas são mais estáveis que os íntrons, que decaem rapidamente após sua remoção dos transcritos primários (ver Cap. 14). Após a marcação e a hibridização ao *chip* de microarranjos, as sequências exônicas apresentam sinais mais intensos do que os íntrons.

Outra característica útil dos microarranjos de genoma inteiro é que eles detectam genes não codificadores, como os que especificam miRNAs. Esses RNAs são geralmente processados a partir de grandes RNAs precursores (pri-RNAs) derivados de unidades de transcrição que possuem 1 a 10 kb de

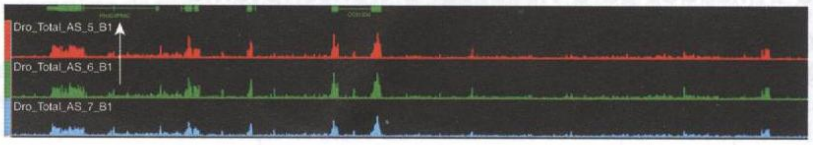


FIGURA 7-23 Microarranjo de genoma inteiro revela detalhes da estrutura de íntron-éxon de um gene. Intervalo de 50 kb no cromossomo 3 de *Drosophila* que contém quatro genes diferentes. A estrutura de íntron-éxon de cada unidade de transcrição é mostrada na parte superior da figura. (Seta branca) A grande região intrônica que pode conter um pequeno ("micro") éxon. RNA total foi extraído de embriões progressivamente mais velhos (vermelho, jovem; verde, mais velho; azul, ainda mais velho) e hibridizado ao microarranjo, composto por sequências de 25 nucleotídeos espaçadas a cada 35 pb ao longo do genoma inteiro. Sinais fortes de hibridização coincidem com os éxons, enquanto há sinais mais fracos nas regiões intrônicas. Com base nos sinais semelhantes em todas as três cores, este gene é expresso em níveis similares em todos os três estágios embrionários testados. (Reproduzida, com permissão, de Manak et al. 2006. *Nat. Genet.* **38**:1151-1158, Fig. 5. © Macmillan.)

comprimento (ver Cap. 20). As unidades de transcrição de pri-RNA são facilmente detectadas por hibridização em microarranjos. Em alguns casos, genes de miRNA contêm íntrons que precisam ser processados antes da produção final do miRNA maduro. Outros tipos de transcritos não codificadores também são detectados, incluindo RNAs "antissenso" dentro de íntrons de genes codificadores de proteínas. É possível que esses RNAs atuem como reguladores para controlar a expressão ou a função de genes codificadores de proteínas.

Os microarranjos levaram a uma observação um tanto surpreendente: cerca de um terço de um genoma típico é transcrito, embora apenas uma fração dessa transcrição corresponda a sequências codificadoras de proteína (apenas 5% no caso do genoma humano). Aparentemente, a maioria da transcrição adicional deve-se a amplas extensões de sequências de DNA intrônico. Vários genes possuem éxons 5′ remotos não codificadores que residem longe (às vezes, 1 megabase ou mais) do corpo principal da sequência codificadora. Em alguns casos, essas regiões intrônicas produzem miRNAs e tipos adicionais de RNAs não codificadores. UTRs (Regiões não traduzidas – *Untranslated regions*) 3′ estendidas representam outra fonte de transcrição não codificadora.

Sequências de DNA reguladoras podem ser identificadas pelo uso de ferramentas de alinhamento especializadas

Tecnologias genômicas são eficazes para identificar genes e determinar a estrutura de suas unidades de transcrição. Uma vez identificadas, uma ampla gama de métodos de bioinformática permite a determinação da estrutura e da função proteica potenciais, por exemplo, se a proteína contém quaisquer domínios ou motivos conhecidos ou se compartilha outras características com proteínas conhecidas. Em especial, o algoritmo BLAST (*Basic Local Alignment Search Tool*) fornece uma abordagem poderosa para buscar, comparar e alinhar sequências de proteínas ou ácidos nucleicos. Buscas com BLAST permitem a rápida comparação de uma determinada sequência exônica com um vasto banco de dados de informações codificadoras de proteínas. Alinhamentos significativos com sequências codificadoras de proteínas de função conhecida (p. ex., proteína de ligação ao DNA, fator de replicação ou receptor de membrana) fornecem pistas imediatas sobre as potenciais atividades do gene e de seus supostos produtos proteicos. Buscas simples com BLAST também podem revelar a identidade de transcritos não codificadores que produzem miRNAs (ver Cap. 20).

Ao contrário de sequências codificadoras de proteínas, a identificação e a caracterização de sequências reguladoras – os trechos do DNA que controlam onde e quando os genes associados estão ligados ou desligados em um organismo – são extremamente desafiadoras, como será visto no Capítulo 19. Na verdade, alguns chamam as sequências reguladoras de "matéria negra" do genoma. Métodos de genoma inteiro estão apenas agora se tornando disponíveis para a identificação desta classe importante de sequência de DNA.

Um subconjunto de sequências reguladoras de vertebrados pode ser identificado usando variações nas buscas de BLAST desenvolvidas para caracterizar sequências codificadoras de proteínas. Reforçadores (ou *enhancers*) célula-específicos contêm sítios de ligação agrupados para uma ou mais proteínas de ligação ao DNA sequência-específicas (ver Cap. 19). Em alguns casos, esse agrupamento é suficiente para a identificação de trechos curtos de alinhamentos de sequências de DNA. Um *software* chamado VISTA alinha as sequências contidas nos genomas de diferentes organismos relacionados em segmentos curtos, de 10 a 20 pb e, assim, identifica sequências não codificadoras imperfeitamente conservadas em trechos de apenas 50 a 75 pb (Fig. 7-24). O peixe baiacu e o camundongo compartilham aproximadamente 10.000 sequências curtas não codificadoras. Imagina-se que muitas dessas sequências correspondam a reforçadores tecido-específicos. Entretanto, é provável que ambos os animais, particularmente os camundongos, tenham pelo menos 100.000 reforçadores. Sendo assim, estes simples alinhamentos de sequências não conseguem capturar a vasta maioria das sequências reguladoras.

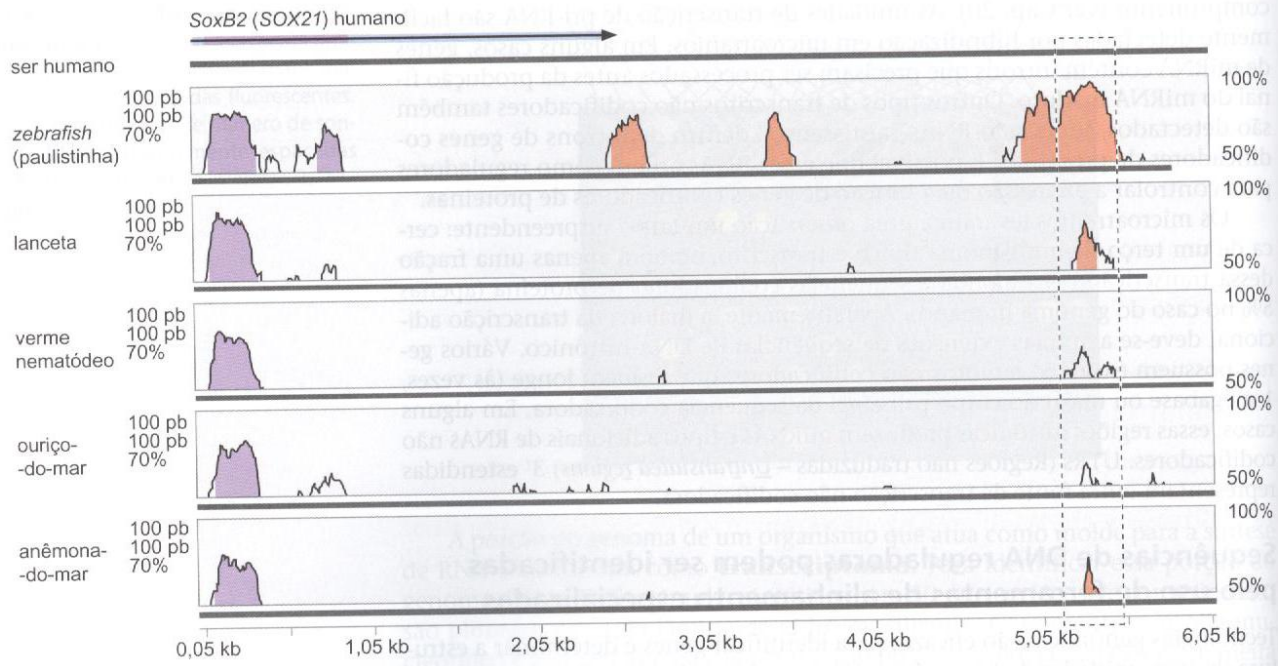


FIGURA 7-24 Comparação do gene *SoxB2* em animais divergentes. Os sinais em lilás correspondem a sequências conservadas na UTR 3′ da unidade de transcrição de SoxB2. Os sinais em cor-de-rosa indicam sequências conservadas que mapeiam a jusante do gene. O retângulo pontilhado identifica reforçadores (*enhancers*) envolvidos na expressão do gene no sistema nervoso. (Adaptada, com permissão, de Royo J.L. et al. 2011. *Proc. Natl. Acad. Sci.* **108:** 14186-14191, Fig. 1A, p. 14187.)

Reforçadores tecido-específicos também podem ser identificados pela triagem de sequências de DNA genômico em busca de potenciais sítios de ligação para proteínas reguladoras conhecidas. Considere-se o caso do gene da α-catenina, que codifica uma molécula de adesão celular. O gene é expresso em vários tecidos diferentes, mas apresenta expressão particularmente forte em células cardíacas precursoras chamadas de cardiomiócitos. Foi possível identificar um reforçador cardíaco-específico pela inspeção de sequências flanqueadoras e intrônicas de α-catenina em busca de sítios de ligação semelhantes aos de proteínas reguladoras conhecidas de células cardíacas, incluindo MEF2C, GATA-4 e E47/HAND (Fig. 7-25). Cada uma dessas proteínas reconhece um espectro de motivos de sequências curtas com 6 a 10 pb. O espectro de sítios de ligação para cada fator é descrito por uma matriz de peso posicional (PWM, *position-weighted matrix*), que pode ser determinada usando diferentes métodos computacionais e experimentais como os ensaios de SELEX (seleção *in vitro*), que serão discutidos em detalhes mais adiante neste capítulo. Quando estas PWMs foram utilizadas para investigar o *locus* de α-catenina, um único agrupamento de supostos sítios para MEF2C, GATA-4 e E47/HAND foi identificado. Estudos experimentais confirmaram que este agrupamento de sítios de ligação, localizado na região flanqueadora 5′ do gene, atua como um autêntico reforçador.

A edição genômica é utilizada para alterar pontualmente genomas complexos

Os métodos anteriores, montagens e anotação genômicas, são descritivos. Eles fornecem atlas detalhados de mapas genômicos inteiros mas não fornecem o tipo de informação funcional que os biólogos moleculares tanto desejam. Entretanto, um método recentemente desenvolvido, a edição genô-

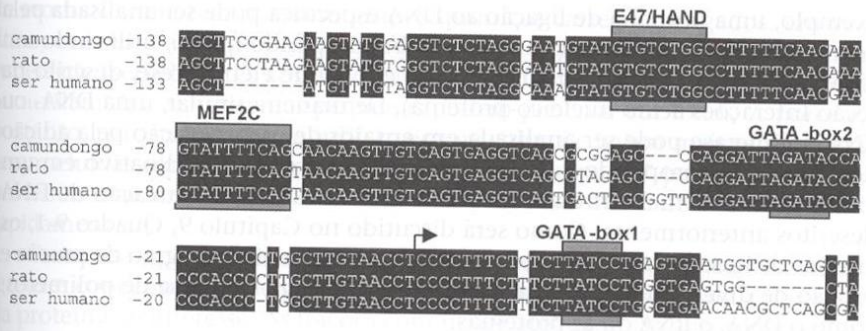


FIGURA 7-25 Identificação de um reforçador cardíaco *in silico*. Uma sequência de ~140 pb na região flanqueadora 5′ do gene da α-catenina é conservada nos genomas do camundongo, do rato e de seres humanos. A sequência conservada contém sítios de ligação para três reguladores cruciais da diferenciação cardíaca: E47/HAND, MEF2C e GATA. A sequência do camundongo atua como um autêntico reforçador cardíaco-específico. A princípio, ele poderia ser identificado por alinhamentos em VISTA (ver Cap. 20, Fig. 20-4) ou pelo agrupamento de proteínas reguladoras cardíacas. (Porção reproduzida, com permissão, de Vanpoucke G. et al. 2004. *Nucleic Acids Res.* **32**: 4155-4165, Fig. 1 © Oxford University Press.)

mica, permite a remoção ou a modificação de segmentos de DNA específicos em um genoma intacto. Essa abordagem envolve a indução de uma quebra de dupla-fita (DSB, *double-strand break*) em um sequência-alvo específica de DNA que estimula a recombinação homóloga para reparar a quebra usando DNA modificado introduzido. Durante o evento de reparo da quebra, alterações desejadas são introduzidas especificamente para modificar a sequência genômica. A clivagem direcionada é realizada por nucleases especialmente customizadas, geralmente nucleases dedo de zinco e "meganucleases", projetadas para clivar em um determinado sítio-alvo no genoma. Entretanto, uma nova classe de nucleases "projetadas" – as nucleases efetoras semelhantes a ativadores transcricionais (TALENs, *transcriptional activator-like effector nucleases*) – apresenta maior eficiência. As TALENs estão emergindo como uma ferramenta importante para a edição genômica direcionada tanto em diferentes organismos-modelo quanto em células-tronco humanas.

PROTEÍNAS

Proteínas específicas podem ser purificadas a partir de extratos celulares

A purificação de proteínas individuais é fundamental para o entendimento de sua função. Embora em alguns momentos a função de uma proteína possa ser estudada em uma mistura complexa, esses estudos podem, com frequência, gerar ambiguidades. Por exemplo, o estudo da atividade de uma DNA-polimerase específica em uma mistura de proteínas totais (como um lisado celular) pode ser mascarado por outras DNA-polimerases e proteínas acessórias parcial ou completamente responsáveis pelas atividades de síntese de DNA observadas. Por essa razão, a purificação de proteínas é uma parte importante nos estudos da sua função.

Cada proteína apresenta propriedades características que tornam a sua purificação relativamente diferente. Isso contrasta com os diferentes DNAs, pois todos apresentam a mesma estrutura helicoidal e são distinguidos somente por sua sequência específica. A purificação de uma proteína procura explorar suas características únicas, incluindo tamanho, carga, formato e, em muitos casos, função.

A purificação de uma proteína requer um protocolo específico

A purificação das proteínas requer uma técnica específica para cada proteína. Na purificação de um DNA, quase sempre a mesma estratégia é utilizada, a hibridização à sua sequência complementar. Como será visto na discussão sobre *imunoblotting*, um anticorpo pode ser utilizado para detectar proteínas específicas da mesma maneira. No entanto, em muitos casos, é mais conveniente utilizar uma medida mais direta para o funcionamento da proteína. Por