Análise da Estrutura e a Expressão de Genes e Genomas

8

CONCEITOS PRINCIPAIS

- Bibliotecas de DNA são coleções de fragmentos de DNA clonados que, coletivamente, representam o genoma de um organismo (bibliotecas de DNA genômico) ou subfrações genômicas.
- Bibliotecas de cDNA são oriundas de RNAs transcritos de modo reverso e, portanto, contêm apenas a parte do genoma que é expressa.
- O sequenciamento de DNA normalmente se baseia na síntese de novas fitas de DNA a partir de modelos de fita simples. Os métodos mais modernos utilizam o sequenciamento em paralelo de um grande número de amostras de DNA, sem utilização de eletroforese.
- O ressequenciamento genômico requer a utilização de uma sequência genômica inicial como referência, visando ao rápido ressequenciamento dos genomas ou de subfrações genômicas de outros indivíduos da mesma espécie.
- Marcadores de DNA são sequências de DNA que possuem uma localização subcromossomal singular, podendo ser, portanto, convenientemente utilizados em ensaios. Mapas de marcadores forneceram importantes mapas esquemáticos para o sequenciamento de genomas complexos, os quais possuem numerosos elementos repetitivos de DNA.
- Marcadores polimórficos de DNA são necessários para construir mapas genéticos. Os diferentes marcadores são genotipados ao longo de diferentes gerações em indivíduos pertencentes a uma linhagem comum.
- Tanto marcadores de DNA não polimórficos como polimórficos podem ser utilizados para construir mapas físicos dos cromossomos, geralmente por meio de painéis de genotipagem de células que possuem diferentes fragmentos de cromossomos individuais.
- Um clone contiguamente ordenado (*contig*) representa uma série de fragmentos de DNA genômicos clonados, dispostos na mesma ordem linear da região subcromossomal a partir da qual foram extraídos. As sobreposições entre cada fragmento de DNA permitem que eles sejam colocados em ordem, fornecendo um importante mapa esquemático para sequenciamento genômico.
- A predição gênica a partir de ferramentas computacionais baseia-se, frequentemente, na identificação de sequências significativamente similares entre uma sequência de DNA (ou oriunda de uma sequência proteica) e a sequência de um gene ou uma proteína conhecidas.
- Os termos *transcriptoma* e *proteoma* descrevem, respectivamente, o grupo completo de transcritos de RNA e de proteínas produzidas por uma célula. Enquanto as diferentes células nucleadas de um organismo possuem genomas estáveis e quase idênticos, transcriptomas e proteomas são dinâmicos, cada um variando de forma bastante significativa entre uma célula e outra.
- Análises da expressão em alta resolução são realizadas com o intuito de obter padrões de expressão gênica detalhados, em células ou tecidos, mas são limitadas à análise de um ou poucos genes por vez.
- Análises em larga escala em paralelo da expressão fornecem informações sobre expressão gênica para milhares de genes por vez, em duas ou mais fontes de células.

O surgimento, nos anos 1970, das tecnologias de clonagem e sequenciamento de DNA inaugurou uma nova era, na qual se tornou possível a caracterização de genes de forma sistemática. Ao sequenciar um gene e o(s) seu(s) transcrito(s), é possível resolver a organização éxon-íntron, bem como predizer a sequência de quaisquer produtos proteicos prováveis.

O progresso inicial foi lento, mas aumentou rapidamente com o aperfeiçoamento tecnológico. No início de 1980, pesquisadores começaram a fazer planos para sequenciar todas as diferentes moléculas de DNA em uma célula (coletivamente conhecida como **genoma**) e, no início dos anos 1990, projetos internacionais sincronizados foram iniciados, visando determinar a sequência completa do genoma humano e de vários organismos-modelo. As sequências genômicas pavimentaram o caminho para análises abrangentes sobre a estrutura e a expressão gênica.

8.1 BIBLIOTECAS DE DNA

As primeiras tentativas para clonagem de fragmentos de DNA humanos em células bacterianas se aproveitaram de transcritos gênicos que eram naturalmente encontrados em altas concentrações em alguns tecidos. Grande parte do mRNA sintetizado em eritrócitos é constituído por mRNA de α- e β-globina, por exemplo. Quando a transcriptase reversa é utilizada para copiar o mRNA de eritrócitos, o cDNA resultante é enriquecido por cDNA de globina, facilitando o seu isolamento.

Para permitir um método de uso mais geral para clonagem de genes e, de fato, de todas as sequências de DNA, as tecnologias subsequentes foram desenvolvidas com o intuito de clonar todas as sequências de DNA constituintes em uma população inicial. As grandes coleções resultantes de clones de DNA ficaram conhecidas como **bibliotecas de DNA**. Embora o PCR possa ser utilizado para gerar bibliotecas de DNA, a clonagem celular de DNA tem sido utilizada, tradicionalmente, em grande parte por ser mais adequada para clonagem de grandes fragmentos de DNA.

Bibliotecas de DNA genômico compreendem cópias fragmentadas de todas as diferentes moléculas de DNA de uma célula

Para qualquer organismo, as células que contêm um núcleo possuem, essencialmente, o mesmo conteúdo de DNA. Para construir uma biblioteca de DNA genômico, o material inicial pode ser o DNA de qualquer célula nucleada representativa, como células brancas do sangue, que são facilmente acessíveis. O DNA genômico isolado é fragmentado normalmente por meio da utilização de endonucleases de restrição que reconhecem uma sequência de 4 pb. Por exemplo, *Mbo*I reconhece a sequência GATC.

Controlando-se a digestão enzimática é possível permitir que a enzima de restrição clive apenas uma porção de todos os sítios de restrição de DNA disponíveis; geralmente a enzima é utilizada em concentrações baixas e o tempo de incubação é curto. Esta *digestão parcial* é utilizada com o objetivo de assegurar que o tamanho médio dos fragmentos produzidos seja um tamanho ótimo para o sistema de clonagem que será utilizado (ver Tabela 6.2).

Além disso, como apenas uma pequena porcentagem dos possíveis sítios de restrição é clivada pela enzima, o DNA é fragmentado *aleatoriamente*. Como o material de início é constituído por milhões de células idênticas, haverá milhões de cópias para cada molécula de DNA diferente. A fragmentação aleatória de DNA significa que, para qualquer que seja a localização em uma molécula de DNA inicial, o padrão de corte irá variar de forma aleatória sobre diferentes cópias daquela molécula de DNA (**Figura 8.1**).

A fragmentação aleatória assegura que a biblioteca conterá tantas representações quantas forem possíveis do DNA inicial. Também haverá clones com insertos sobrepostos (ver Figura 8.1). Isso significa que qualquer clone de DNA selecionado a partir da biblioteca poderá ser utilizado para recuperar clones com insertos que se sobrepõem. Como resultado, os clones podem ser arranjados em uma ordem que corresponda à ordem cromossomal das sequências de DNA clonadas. Isso será discutido posteriormente no contexto dos projetos genoma, na Seção 8.3.

Bibliotecas de cDNA compreendem cópias de DNA de diferentes moléculas de RNA em uma célula

Ao contrário do DNA genômico, o conteúdo de RNA das células de um indivíduo pode variar enormemente de um tipo celular para outro, bem como de um estágio do desenvolvimento para outro. A amostra inicial é constituída, em geral, pelo RNA total de um tecido, de uma linhagem celular ou de um estágio específico do desenvolvimento embrionário.

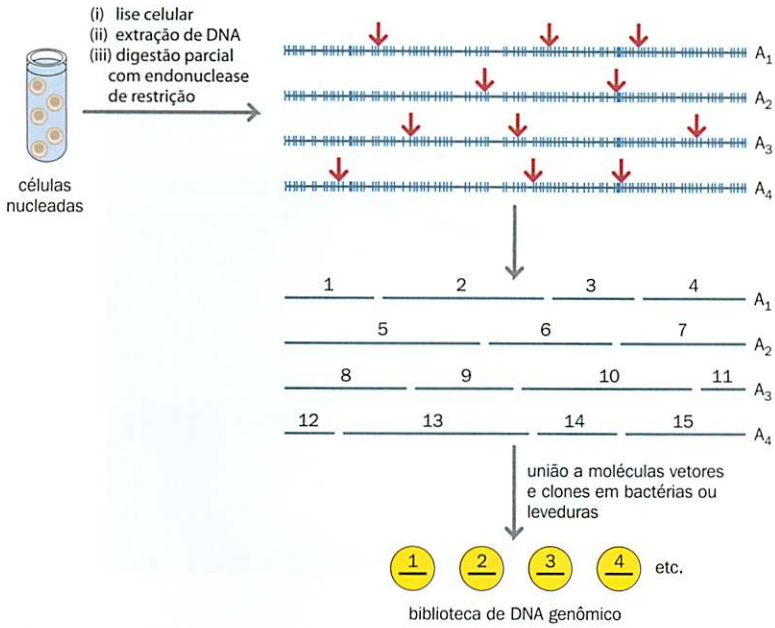

Figura 8.1 Construindo uma biblioteca de DNA genômico. Todas as células nucleadas de um indivíduo terão, essencialmente, o mesmo conteúdo de DNA genômico; assim, qualquer célula que seja facilmente acessível (como células sanguíneas) pode ser utilizada como fonte de material. Como o DNA é extraído a partir de um grande número de células com moléculas de DNA idênticas, o DNA isolado irá conter um grande número de sequências de DNA idênticas, como A_1, A_2, A_3 e A_4, aqui ilustrados. Entretanto, a digestão parcial com uma endonuclease de restrição irá clivar o DNA em apenas uma fração de sítios de restrição disponíveis (pequenas barras azuis verticais indicam todos os sítios de restrição que a enzima poderá clivar, se utilizada sob condições normais; flechas verticais vermelhas marcam as posições do pequeno número de sítios de restrição que são clivados). Como resultado, o padrão irá diferir entre cópias idênticas das mesmas moléculas de DNA cromossomais, quase resultando em uma clivagem aleatória. Isso irá gerar uma série de fragmentos de restrição que, se forem oriundos da mesma região subcromossomal, podem compartilhar alguma sequência de DNA em comum (p. ex., fragmento 6 da molécula original A_2 sobrepõe-se parcialmente aos fragmentos 2 e 3 de A_1, fragmentos 9 e 10 de A_3 e fragmentos 13 e 14 de A_4).

Até recentemente, a fração de mRNA codificante para um determinado polipeptídeo era considerada praticamente a única classe de RNA de interesse, e, como quase todos os mRNA são poliadenilados, mRNAs com poli(A)$^+$ seriam selecionados a partir da ligação específica a uma sequência oligo(dT) ou poli(U) complementar aderida a uma matriz sólida. A transcriptase reversa seria então utilizada para converter os mRNAs poli(A)$^+$ isolados em uma cópia de cDNA dupla-fita (**Figura 8.2**). Atualmente, ficou claro que uma fração substancial de RNA funcional não é poliadenilada. Assim, o RNA celular total é preparado e transformado em cDNA a partir da utilização de *primers* (iniciadores) hexanucleotídicos randômicos em conjunto com a transcriptase reversa.

Uma abordagem inicial lógica para construção de bibliotecas de DNA foi fornecer uma rota rápida em favor da clonagem de genes (sequências codificantes de DNA humano correspondem a apenas cerca de 1% do DNA de uma célula humana). Devido às diferenças nos padrões de expressão gênica, uma grande variedade de bibliotecas de cDNA tem sido construída a partir de diferentes tecidos, linhagens celulares e estágios do desenvolvimento, para diversos organismos.

Para serem úteis, as bibliotecas de DNA precisam ser convenientemente selecionadas e disseminadas

Assim como qualquer outro sistema de clonagem de DNA baseado em células, a triagem de bibliotecas de DNA requer a seleção inicial de células transformadas, normalmente por meio da utilização de moléculas vetores que conferem resistência a antibióticos. Como descrito na Seção 6.1, recombinantes também são frequentemente identificados pela inativação insercional; o sítio de inserção é projetado para interromper um gene marcador dentro de um vetor de clonagem; assim, qualquer que seja o tamanho do inserto, este permitirá que o gene marcador seja inativado.

Seleção da biblioteca

Para identificar um clone contendo um DNA recombinante específico, a biblioteca é selecionada a partir de dois métodos. Na **hibridização de colônias**, alíquotas de clones celulares

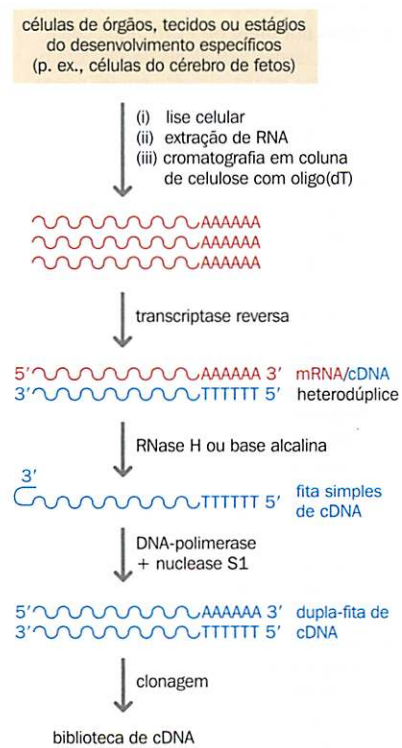

Figura 8.2 Construindo uma biblioteca de cDNA. Como mostrado aqui, a transcriptase reversa utiliza, frequentemente, um *primer* oligo(dT) para iniciar a síntese da fita de cDNA. Mais recentemente, misturas de *primers* hexanucleotídicos randômicos têm sido utilizadas, fornecendo uma representação mais normal de sequências. A RNase H irá digerir especificamente o RNA que estiver ligado ao DNA em um híbrido RNA-DNA. A extremidade 3′ do cDNA fita simples resultante possui uma tendência de virar para trás, formando uma pequena estrutura em forma de grampo (*hairpin*). Isto pode ser utilizado para iniciar a síntese da segunda fita pela DNA-polimerase, sendo que a pequena alça (*loop*) resultante, conectando as duas fitas, pode então ser clivada pela nuclease S1, a qual cliva especificamente regiões do DNA que são fita simples.

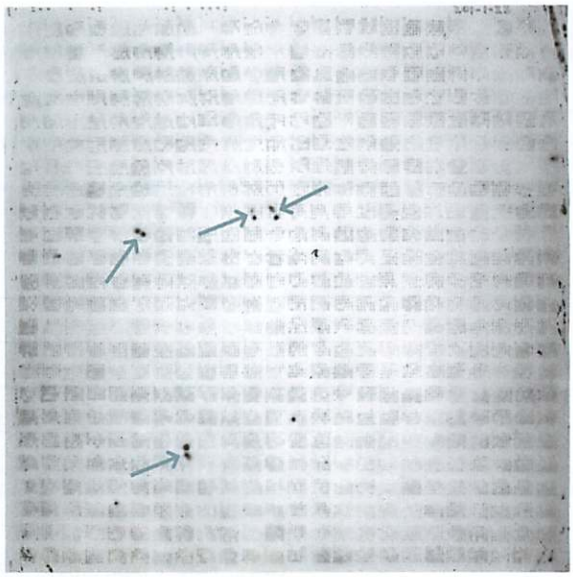

Figura 8.3 Selecionando uma biblioteca de DNA por meio da hibridização de colônias. A autorradiografia é de uma membrana contendo 17.664 clones YAC humanos (DNA total de clones de leveduras individuais contendo cromossomos artificiais de levedura humanos) gradeados em 6 × 6 unidades clonais. Os sinais de hibridização incluem sinais fracos obtidos com uma sonda marcada com S^{35} de todos os clones de DNA total de levedura mais os sinais de hibridização fortes obtidos com uma sonda específica marcada com P^{32} do cromossomo X humano (*DXYS646*). [Cortesia de Mark Ross, Instituto Sanger, de Ross MT & Stanton VP (1995) *Current Protocols in Human Genetics* vol. 1. Com permissão de Wiley-Liss, Inc., filial de John Wiley & Sons, Inc.]

oriundos da biblioteca de DNA são dispostos pontualmente em forma de uma grade de alta densidade sobre uma membrana de nitrocelulose ou de náilon e submetidos à lise celular e à desnaturação do DNA (ver Figura 7.13). A adição de uma sonda de hibridização de interesse marcada pode, então, identificar células sobre a membrana que possuem o inserto de DNA relacionado (**Figura 8.3**), permitindo que o clone parental seja identificado. De forma alternativa, é possível realizar a triagem de bibliotecas com o uso de PCR, por meio da seleção de diferentes combinações de clones a partir de placas de microtitulação (**Figura 8.4**).

Amplificação da biblioteca e disseminação

Bibliotecas recém-sintetizadas são ditas não amplificadas, embora as células transformadas inicialmente tenham sido amplificadas para formar colônias celulares visíveis. Para propagar as bibliotecas, colônias individuais são colhidas e pinçadas no formato de uma grade sobre membranas adequadas ou dentro de poços de placas de microtitulação, onde elas podem ser estocadas por longos períodos a −70°C na presença de um agente estabilizador de célula, como o glicerol.

Algumas bibliotecas estão em alta demanda e são, portanto, distribuídas em larga escala. Para isso, a biblioteca original precisa ser amplificada. Para transporte e estoque a longo prazo, pequenas alíquotas das bibliotecas originais são diluídas em meio celular

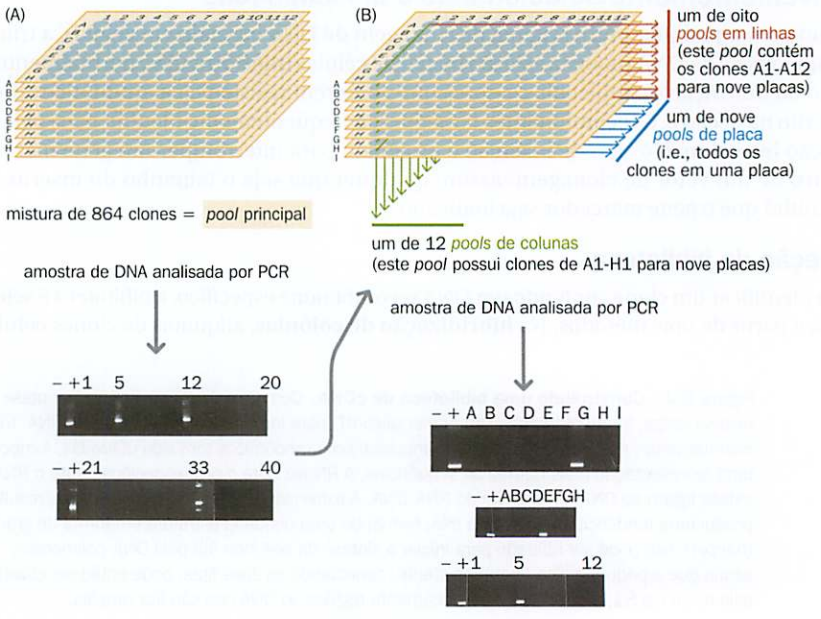

Figura 8.4 Seleção de uma biblioteca de cDNA por meio de PCR. Aqui, cerca de 35 mil clones individuais de YAC foram depositados individualmente em 96 poços de 360 microplacas de titulação. Para facilitar a triagem, um total de 40 *pools* principais foi gerado a partir da combinação de todos os 864 clones, em grupos de nove microplacas de titulação (placas A-I). (A) Triagem primária. Esta parte envolve a análise por PCR de 40 *pools* principais. Neste exemplo, três *pools* principais foram positivos quando referenciados contra controles positivo (+) e negativo (−): *pools* 5, 12 e 33. (B) Triagem secundária. Aqui, YACs individuais foram identificados a partir do ensaio de diferentes subgrupos dos 864 YACs em um *pool* principal positivo, neste caso, o *pool* principal de número 12. A triagem tridimensional de cada um dos *pools* das nove placas (96 YACs cada), oito *pools* em linha (106 YACs cada) e 12 *pools* em coluna (72 YACs cada), identificou um YAC positivo na placa 12G (painel superior), linha E (painel do meio), coluna 5 (painel inferior). [Cortesia de Sandie Jones, Universidade de Newcastle e adaptada de Jones MH, Khwaja OS, Briggs H et al. (1994) *Genomics* 24, 266-275. Com permissão de Academic Press Inc.]

contendo agente estabilizador de células – as bibliotecas podem ser posteriormente regeneradas por meio do plaqueamento, permitindo o crescimento de todas as células que contêm DNA recombinante.

Durante o estágio de amplificação, no entanto, diferentes colônias podem crescer com diferentes taxas. Por exemplo, o crescimento de células específicas pode ser retardado caso elas contenham um DNA recombinante que afete o seu metabolismo de alguma maneira. A amplificação pode, portanto, resultar em uma distorção da representação original dos clones celulares da biblioteca não amplificada.

8.2 SEQUENCIANDO O DNA

O desenvolvimento de bibliotecas de DNA tornou possível, em princípio, a realização abrangente de sequenciamento de DNA. Até recentemente, o sequenciamento moderno de DNA era baseado em um método de sequenciamento enzimático, desenvolvido pela primeira vez na década de 1970, onde uma DNA-polimerase era utilizada para sintetizar novas cadeias de DNA a partir de um modelo de DNA fita simples clonado, constituído por milhões de cópias idênticas de uma sequência específica de DNA. Inicialmente, o sequenciamento de DNA era lento e demorado, além de os dados de saída serem limitados, mas a necessidade de evoluir para o sequenciamento em larga escala de clones de bibliotecas levou a avanços tecnológicos.

O sequenciamento dideoxi de DNA envolve a síntese enzimática de DNA utilizando terminadores de cadeia base-específicos

O método de sequenciamento de DNA que foi utilizado para sequenciar o genoma humano, assim como muitos outros genomas, era um método enzimático de sequenciamento desenvolvido de modo pioneiro por Fred Sanger, em meados de 1970. Esse método se baseia na inibição randômica do alongamento da cadeia, criando fitas de DNA recém-sintetizadas de vários comprimentos, podendo ser separadas de acordo com o tamanho. O DNA precisa estar na disposição de fita simples, agindo como um modelo para construção de uma nova fita de DNA complementar *in vitro*, por meio da utilização de uma DNA-polimerase adequada.

O substrato para o sequenciamento do DNA era, frequentemente, um DNA recombinante que seria desnaturado, permitindo que um *primer* de sequenciamento específico pudesse ser utilizado para direcionar a síntese da nova fita; ou os fragmentos de DNA seriam clonados em fagomídeos que eram manipulados para produzir um DNA recombinante fita simples. Uma alternativa que está se tornando bastante comum, é utilizar o DNA produzido a partir da amplificação por PCR e convertê-lo para a forma fita simples para que aja como um modelo do sequenciamento. O produto final de qualquer um dos métodos é uma população de várias cópias idênticas do DNA que será sequenciado.

O sequenciamento é conduzido em quatro reações paralelas, cada uma contendo os quatro dNTPs (dATP, dCTP, dGTP e dTTP) além de uma pequena proporção de um dos quatro análogos de dideoxinucleotídeos (ddNTPs) que servirão como terminadores de cadeia base-específicos. Um ddNTP é muito parecido com o seu homólogo de dNTP, porém não possui o grupo hidroxil na posição do carbono 3′ e também na posição do carbono 2′ (**Figura 8.5**). O ddNTP pode ser incorporado na cadeia crescente de DNA formando uma ponte fosfodiéster entre o seu carbono 5′ e o carbono 3′ do nucleotídeo recém-incorporado. Entretanto, como ele não possui o grupo hidroxila da posição 3′, qualquer dideoxinucleotídeo que for incorporado na cadeia crescente de DNA não poderá participar da ligação fosfosdiéster no seu carbono 3′. Uma vez que o dideoxinucleotídeo tenha sido incorporado, portanto, causará a terminação abrupta do alongamento da cadeia.

Assegurando que um dos quatro dNTPs, ou que o *primer*, esteja marcado, a cadeia crescente de DNA se tornará marcada também. Ajustando-se a concentração de dideoxinucleotídeos para que seja muito mais baixa do que os deoxinucleotídeos análogos correspondentes, haverá competição entre um dideoxinucleotídeo específico e o seu deoxinucleotídeo análogo para inclusão na cadeia crescente de DNA. O deoxinucleotídeo está presente em excesso; quando incorporado, o alongamento da cadeia continua, mas ocasionalmente o dideoxinucleotídeo será incorporado na cadeia crescente, finalizando a polimerização e, assim, causando a sua terminação. Cada reação é, portanto, uma reação *parcial*, pois a terminação da cadeia ocorrerá randomicamente em uma das possíveis escolhas para um tipo específico de base em qualquer fita de DNA.

Como a amostra de DNA representa uma população de moléculas idênticas, cada uma das quatro reações base-específicas gerará uma coleção de fragmentos de DNA marcados, com diferentes comprimentos. Cada um dos fragmentos em uma reação terá em

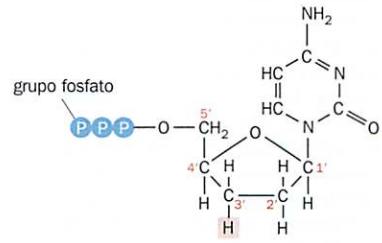

Figura 8.5 Estrutura de um dideoxinucleotídeo: 2′, 3′-dideoxi CTP (ddCTP). O grupo hidroxila ancorado ao carbono 3′ em nucleotídeos normais é trocado por um átomo de hidrogênio (sombreado).

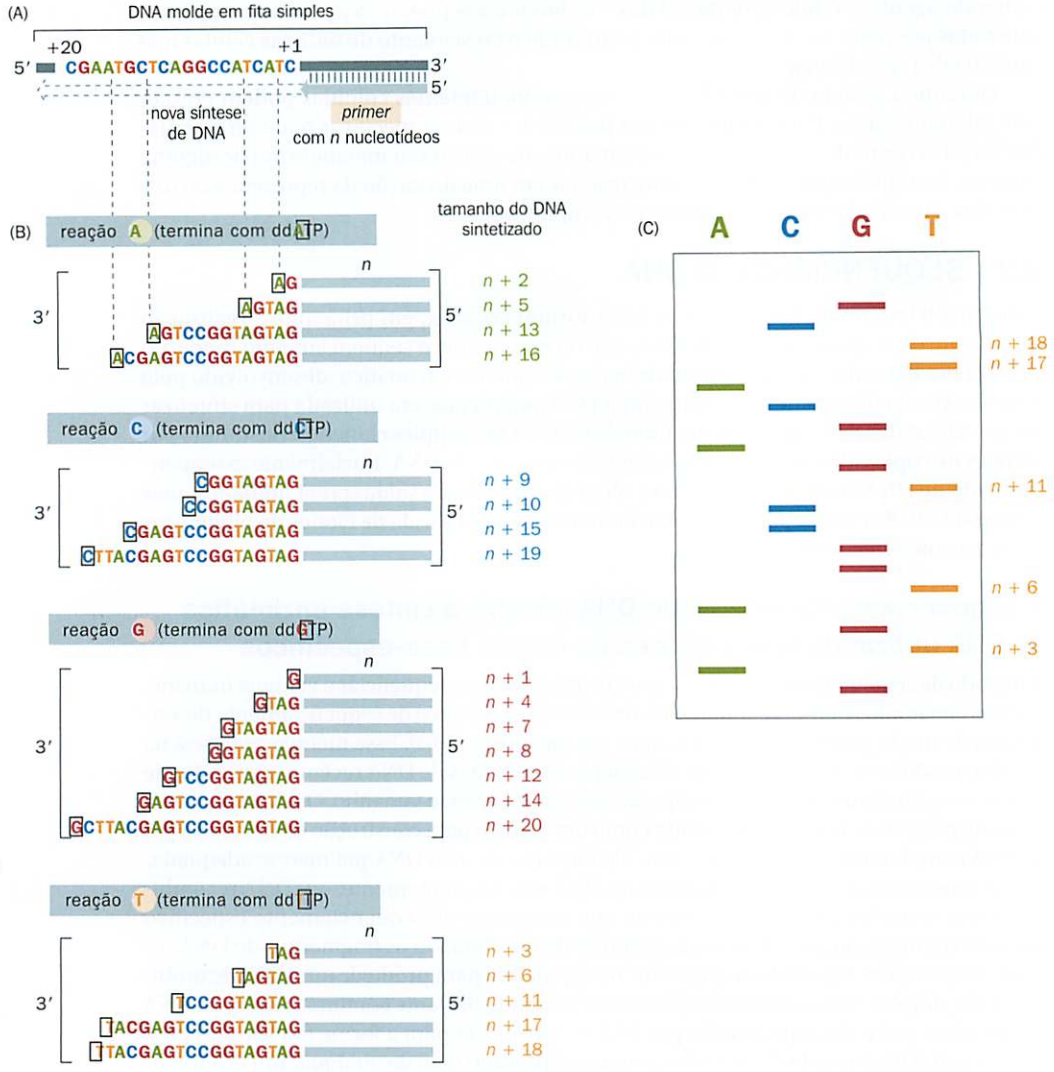

Figura 8.6 Sequenciamento de DNA pelo método dideoxi. (A) Com a utilização de um *primer* de cerca de 20 nucleotídeos, uma sequência complementar é sintetizada a partir de um molde de DNA fita simples. (B) Quatro reações em paralelo base-específicas são realizadas, cada uma com todos dNTPs mais um ddNTP que compete com o dNTP homólogo pela inserção na cadeia crescente. Por exemplo, a reação A utiliza dATP, dCTP, dGTP, dTTP e também ddATP, o qual compete com dATP causando a terminação da cadeia quando ele é incorporado. Como há muitas cópias idênticas ao molde, uma série de diferentes fragmentos é produzida, dependendo do ponto no qual o ddATP foi inserido. Os fragmentos terão uma extremidade 5′ em comum (definida pela sequência de *primer*) mas extremidades 3′ variáveis, dependendo de onde o ddA (mostrado como um A enclausurado dentro de uma caixa) tenha sido inserido. Por exemplo, na cadeia de reação específica para A, a extensão ocorre até que um ddA seja incorporado. Isso levará a uma população de fragmentos de DNA com nucleotídeos variando em tamanho de $n+2$, $n+5$, $n+13$, $n+16$, etc. (C) O fracionamento por tamanho em gel de poliacrilamida permite que a sequência seja desvendada, sendo que a sequência de $n+1$ a $n+20$ será GATGATGGCCTGAGCATTCG, a sequência complementar reversa da mostrada em (A).

comum a extremidade 5′ (definida pelo *primer* do sequenciamento). Entretanto, as extremidades 3′ disponíveis são variáveis, pois a inserção do dideoxinucleotídeo selecionado ocorre aleatoriamente em uma das várias posições diferentes que aceitarão a base-específica (**Figura 8.6**).

Fragmentos que diferem em tamanho por até mesmo um único nucleotídeo podem ser separados de acordo com o tamanho em um gel de poliacrilamida, o qual contém uma alta concentração de ureia ou outros agentes desnaturantes, permitindo assim que o DNA permaneça fita simples ao longo da migração.

A automatização do sequenciamento dideoxi aumentou a sua eficiência

Máquinas de sequenciamento automatizado do DNA que utilizavam marcação fluorescente do DNA foram desenvolvidas no início da década de 1990. Quatro diferentes marcadores fluorescentes são utilizados nas quatro reações base-específicas. Por meio da seleção de marcadores com emissão de diferentes comprimentos de onda, todas as quatro reações puderam ser carregadas em uma única amostra do gel. Ao longo da eletroforese, os fragmentos de DNA são submetidos a uma fonte de excitação, como um *laser*, enquanto um monitor detecta e grava o sinal de fluorescência, conforme o DNA passa através de um ponto fixo no gel (**Figura 8.7**). Isso permite que os resultados estejam dispostos na forma de perfis que variam de acordo com a intensidade para cada um dos fluoróforos coloridos, ao mesmo tempo que a informação é estocada eletronicamente.

Os sequenciadores de DNA automatizados mais antigos usavam géis de poliacrilamida sobre uma placa, mas, com o aumento da capacidade de sequenciamento de DNA, tornou-se

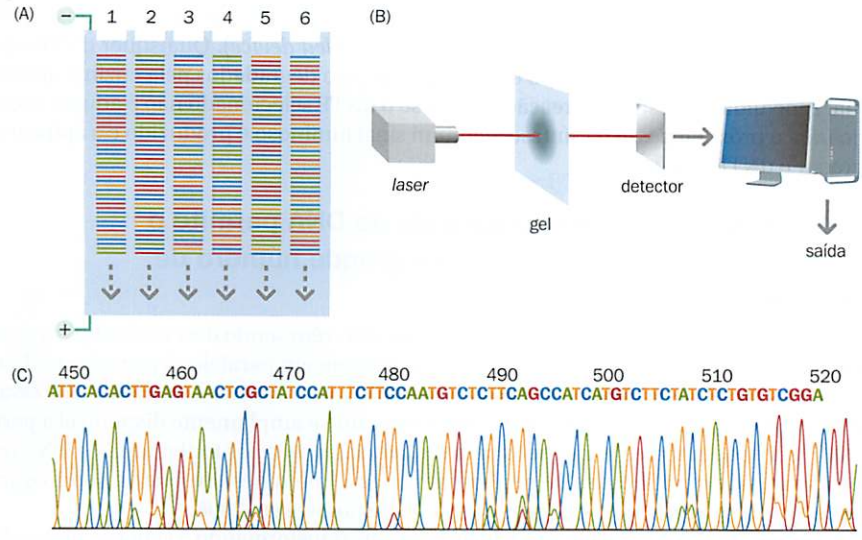

Figura 8.7 Sequenciamento de DNA automatizado com *primers* fluorescentes. (A) Quatro corantes fluorescentes separados são utilizados como marcadores para as reações base-específicas (o marcador pode ser incorporado por meio da sua ligação a um ddNTP base-específico ou a um *primer*, tendo quatro grupos de *primers* diferentes correspondentes às quatro reações). Amostras contendo misturas de todas as quatro reações base-específicas são fracionadas de acordo com o tamanho por eletroforese em gel de poliacrilamida (aqui, os fragmentos são mostrados migrando para baixo através de uma placa de gel). (B) Enquanto os fragmentos migram durante a eletroforese, um feixe de *laser* é focado em uma posição específica do gel. Conforme os fragmentos de DNA individuais passam por essa posição, o *laser* promove a fluorescência do marcador. A fluorescência máxima ocorre em diferentes comprimentos de onda para os quatro marcadores; a informação é gravada eletronicamente e a sequência interpretada é estocada em um banco de dados computacional. (C) Exemplo da saída de uma sequência de DNA mostrando uma sucessão de perfis de intensidade específicos de marcadores (e, portanto, base-específico). O exemplo ilustrado mostra uma sequência de cDNA do gene poli-homeótico humano *PHC3*.

possível o sequenciamento capilar. Nessa técnica, as amostras de DNA migram através de um tubo capilar de vidro, longo e extremamente fino (diâmetro de 0,1 mm), contendo gel de poliacrilamida. Assim como a tecnologia de géis sobre placas, a tecnologia capilar lê a base da sequência conforme o DNA migra ao longo do gel, mas um grau de automação muito mais elevado pode ser obtido, evitando-se a necessidade de uma enorme placa molde de gel.

O pirossequenciamento iterativo grava as sequências de DNA enquanto as moléculas de DNA estão sendo sintetizadas

Com pequenas modificações, o método de sequenciamento dideoxi sustentou a genética molecular por três décadas. Esse método, no entanto, possui uma desvantagem, pois se baseia na eletroforese em gel para separar segmentos de DNA recém-sintetizados. Isso não apenas permite que o método seja lento, porém, mais importante que isso, torna difícil o sequenciamento de um grande número de fragmentos de DNA por vez – os equipamentos mais avançados de sequenciamento de DNA pelo método dideoxi possuem a capacidade de sequenciar no máximo 96 amostras por vez. Assim, o resultado do sequenciamento fica limitado a sequências de 30 a 60 kb a cada 3 a 4 horas por rodada de eletroforese.

Tecnologias fundamentalmente diferentes de sequenciamento de DNA que não utilizavam eletroforese em gel não foram desenvolvidas até início e meados da década de 2000. Um importante avanço foi o desenvolvimento de métodos com a capacidade de gravar a sequência de DNA enquanto a fita de DNA era sintetizada pela DNA-polimerase a partir do molde fita simples. Ou seja, o método de sequenciamento foi capaz de monitorar a incorporação de cada nucleotídeo na cadeia de DNA crescente e de identificar qual nucleotídeo estava sendo incorporado a cada passo.

No centro da primeira abordagem desse tipo estava um método conhecido como pirossequenciamento, o qual foi desenvolvido, inicialmente, para avaliar polimorfismos pontuais de nucleotídeos. Para utilização no sequenciamento, sucessivas reações de pirossequenciamento (*pirossequenciamento iterativo*) são realizadas para gravar a sequência de DNA conforme ela é sintetizada. As cadeias de DNA são sintetizadas a partir de precursores de dNTP (deoxinucleosídeo trifosfatado), e a DNA-polimerase promove, naturalmente, a clivagem entre os fosfatos α e β, incorporando um dNMP (contendo o fosfato α) ao DNA, deixando para trás um resíduo de pirofosfato (PPi) composto pelos fosfatos β e γ. O pirossequenciamento explora o fato da liberação de um pirofosfato cada vez que um nucleotídeo é incorporado com sucesso em uma cadeia de DNA crescente.

Reações enzimáticas sequenciais são utilizadas para detectar a liberação do PPi. Primeiramente, a ATP sulforilase converte de modo quantitativo o PPi em ATP na presença de adenosina 5′-fosfossulfato. A seguir, a liberação de ATP promove uma reação onde a luciferase converte a luciferina em oxiluciferina, um produto que gera luz visível em quantidades que são proporcionais à quantidade de ATP. Assim, cada vez que um nucleotídeo é incorporado, um sinal de luz é detectado.

No pirossequenciamento iterativo, os dNTPs individuais são fornecidos sequencialmente (ao contrário do sequenciamento dideoxi, onde uma mistura de dNTPs é utilizada). Se o dNTP selecionado é aquele que pode fornecer o dNMP requerido para a continuação

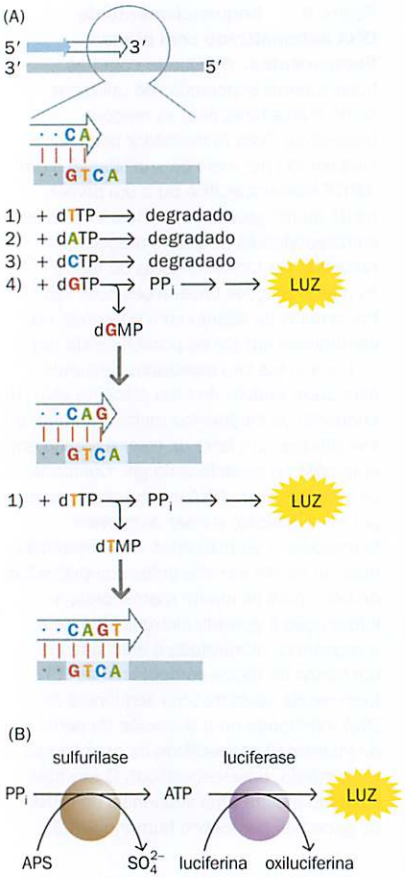

Figura 8.8 O princípio por trás do pirossequenciamento. (A) Inserção de bases no pirossequenciamento. No pirossequenciamento, uma DNA-polimerase sintetiza uma cadeia de DNA utilizando um molde de DNA fita simples e os quatro dNTPs normais. Em vez de ter um mistura dos quatro dNTPs, os dNTPs individuais são fornecidos sequencialmente. Quando o dNTP correto é fornecido, a incorporação do nucleotídeo dNMP é rastreada por meio da produção simultânea de um grupo pirofosfato (PP_i), o qual é utilizado para gerar um sinal luminoso. Se um dNTP incorreto é fornecido, ele é degradado pela enzima apirase. Neste exemplo, o primeiro dNMP a ser incorporado é o G, como indicado pela produção de luz em uma reação específica para G, seguida pela reação T. (B) A inserção da base correta é monitorada por meio da produção de luz em uma reação de dois passos. O PP_i liberado é utilizado pela enzima ATP sulfurilase para gerar ATP a qual, subsequentemente, leva a reação com luciferase a produzir luz, detectada por uma câmera com dispositivo de carga acoplada (CCD).

do alongamento da cadeia, então o PPi é liberado, produzindo um sinal luminoso que é gravado por uma câmera CCD (do inglês *charge-coupled device*). Quaisquer dNTPs que não forem utilizados, assim como o excesso de ATP, são degradados pela enzima apirase, a qual é incluída na mistura da reação. Assim, se o dNTP selecionado não é aquele necessário para o próximo passo da síntese, nenhum sinal luminoso é produzido e a apirase irá degradar o dNTP (**Figura 8.8**).

O sequenciamento massivo paralelo de DNA permite o sequenciamento simultâneo de um grande número de diferentes fragmentos de DNA

Vários métodos massivos de sequenciamento paralelo vêm sendo desenvolvidos, os quais podem executar bilhões de reações de sequenciamento em paralelo. A primeira onda de tecnologias de sequenciamento de próxima geração que foram desenvolvidas utilizava o método de PCR para amplificar o DNA-alvo, tornando-se amplamente disponível a partir de 2007. Subsequentemente, métodos para sequenciar moléculas individuais de DNA que não haviam sido amplificadas de nenhuma maneira foram desenvolvidos, sendo que a primeira máquina para esse propósito se tornou disponível em 2008.

O massivo sequenciamento paralelo de DNA está transformando a genética molecular. Tais métodos estão sendo extensivamente utilizados no rápido ressequenciamento do genoma para vários propósitos, inclusive no sequenciamento de genoma pessoal e na identificação de mutações. Além disso, esses métodos estão permitindo uma análise mais abrangente do **transcriptoma**, a coleção das diferentes moléculas de RNA nas células. Ainda, como será visto no Capítulo 12, eles fornecem resoluções rápidas para várias análises que procuram identificar interações proteína-DNA e sítios de ligação para proteínas regulatórias.

Sequenciamento massivo paralelo de DNA amplificado

Os primeiros métodos de sequenciamento massivo em paralelo a serem desenvolvidos utilizavam a amplificação por PCR e envolviam o sequenciamento à medida que ocorria a síntese. O pirossequenciamento massivo em paralelo foi desenvolvido nos anos 2000 pela companhia 454 Life Sciences e subsequentemente comercializado pela empresa Roche. O método envolve a quebra do DNA presente na amostra em pequenos fragmentos (300-500 pb) e a preparação de moldes fita simples. Dois tipos diferentes de oligonucleotídeos adaptadores são ligados às extremidades dos fragmentos de DNA para fornecer extremidades iniciadoras universais para amplificação. Após a desnaturação, os pequenos fragmentos fita simples que possuem os diferentes adaptadores ligados a extremidades diferentes são selecionados. O próximo passo é unir moléculas de DNA a esferas (**Figura 8.9A**), obtendo vantagem da ligação biotina-estreptavidina. Como descrito na Seção 7.2, a proteína bacteriana estreptavidina possui uma extraordinária afinidade de ligação pela vitamina biotina; sendo assim, ao desenvolver um dos adaptadores para que o mesmo possua uma biotina, as moléculas de DNA se ligarão a esferas recobertas com a parceira de ligação da biotina, ou seja, a estreptavidina.

Moldes de DNA fita simples são imobilizados em esferas, e estas são separadas umas das outras por meio de uma emulsão óleo-água, onde cada gotícula contém uma única esfera, assim como os reagentes necessários para o PCR. Essas gotículas são conhecidas como microrreatores (**Figura 8.9B**). Após a amplificação por PCR, há cerca de 10 milhões de cópias de um fragmento de DNA imobilizadas em uma esfera. A emulsão é então quebrada e as esferas são carregadas em poços, com capacidade de picolitros, sobre uma placa (uma esfera por poço), as quais são então cobertas com pequenas esferas que possuem ATP sulfurilase e luciferase aderidas a sua superfície (ver Figura 8.9B). Uma sequência fixa de precursores de dNTP (primeiro T, depois A, então C e, por último, G) é colocada sobre as esferas e o sinal luminescente é emitido cada vez que um nucleotídeo é incorporado; isto é gravado para cada posição individual.

O método de pirossequenciamento massivo em paralelo desenvolvido pela companhia 454 Life Sciences (e comercializada pela empresa Roche) produz leituras de sequências razoavelmente longas de DNA, mas são centenas de milhares de sequências que podem ser lidas em paralelo. Assim, o rendimento é quase 10 milhões de vezes superior ao do sequenciamento pelo método dideoxi (**Tabela 8.1**). A tecnologia de sequenciamento relacionada, a Solexa, foi desenvolvida mais recentemente pela empresa Illumina e possui um rendimento de sequenciamento ainda maior, embora seja acompanhada por sequências individuais menores.

Uma tecnologia alternativa, o sequenciamento massivo em paralelo baseado em ligação, foi desenvolvida pela companhia Applied Biosystems. A sua tecnologia SOLiD™

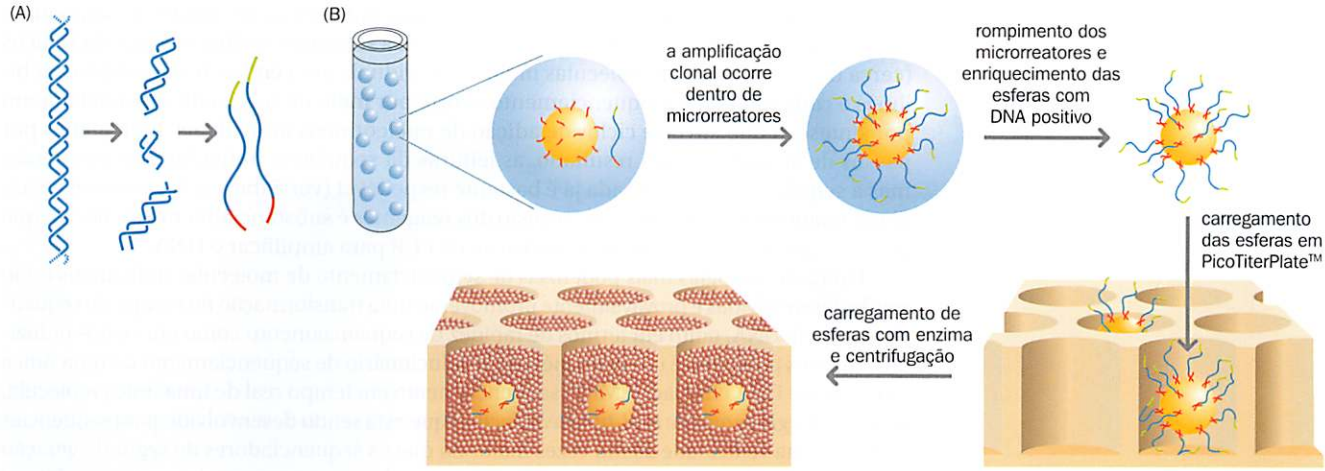

Figura 8.9 Preparação da amostra para pirossequenciamento massivo em paralelo. (A) O DNA genômico é isolado, fragmentado e ligado a adaptadores oligonucleotídicos e separados em fita simples. (B) Os fragmentos são ligados a esferas sob condições que favorecem um fragmento por esfera. Após, as esferas são capturadas em gotas na forma de um microrreator (uma emulsão contendo uma reação de PCR misturada em óleo). A amplificação clonal por PCR ocorre dentro de cada gota, resultando em esferas onde cada uma carrega 10 milhões de cópias de um molde de DNA único. A emulsão é rompida para liberar as esferas e as fitas de DNA são desnaturadas. Esferas únicas carregando clones de DNA fita simples são depositadas em poços individuais em uma placa de fibra óptica (PicoTiterPlate™) que contém 1,6 milhões de poços. Cada poço irá conter um picolitro de reação de sequenciamento. Esferas menores carregando enzimas imobilizadas requeridas para o pirossequenciamento (Figura 8.8B) são depositadas em cada poço para permitir que a reação de pirossequenciamento ocorra. [Adaptada de Margulies M, Egholm M, Altman WE et al. (2005) *Nature* 437, 376-380. Com permissão de Macmillan Publishers Ltd.]

(sequenciamento por meio da detecção da ligação de oligonucleotídeos, do inglês *Sequencing by Oligonucleotide Ligation Detection*) também utiliza uma estratégia de PCR baseada em emulsão para amplificar fragmentos de DNA individuais. Os produtos da amplificação são, então, depositados de forma randômica em uma determinada ordem com sondas de oligonucleotídeos marcados com fluorescência. O método SOLiD também possui um alto rendimento de sequenciamento mas, comparativamente, com pequenas sequências individuais.

Sequenciamento de uma única molécula

Uma nova onda de tecnologias, algumas vezes referida como sequenciamento de terceira geração, permite o sequenciamento de moléculas individuais de DNA, as quais não são amplificadas de nenhuma maneira. Erros no sequenciamento que surgem ocasionalmente como resultado da amplificação do molde de DNA podem ser evitados ou minimizados, e, ademais, essas tecnologias são mais simples do que aquelas que utilizam moldes amplificados, além de possuírem um custo projetado de sequenciamento muito mais baixo. A primeira tecnologia de sequenciamento de moléculas individuais a ser comercializada foi desenvolvida pela Helicos Biosciences, cuja HeliScope™ se tornou disponível em 2008.

TABELA 8.1 Exemplos de metodologias de massivo sequenciamento paralelo de DNA

Metodologia	Plataforma de sequenciamento de DNA	Tamanho da leitura por reação (nucleotídeos)	Saída	Comentários
Sequenciamento por síntese de DNA amplificado por PCR	Sequenciador Roche GS FLX	>300	>0,45 Gb por rodada de 7 horas	Tamanho das leituras é grande, mas os custos dos reagentes são altos e a saída, limitada
	Illumina Genome Analyzer IIx	~100	18 Gb por rodada de 4 dias	Atualmente é a mais utilizada, mas os custos com reagentes são altos
		2 ×~100[a]	35 Gb por rodada de 9 dias	
Sequenciamento por ligação de DNA amplificado por PCR	Applied Biosystems SOLiD 3	~50	Até 30 Gb por rodada de 7 dias	Tamanho das leituras é pequeno, cada rodada leva tempo e tem altos custos com reagentes
		2 ×~50[a]	Até 50 Gb por rodada de 14 dias	
Sequenciamento de molécula individual	Helicos Biosciences HeliScope	~30	~40 Gb por rodada de 8 dias	Baixo custo de reagentes, mas o tamanho das leituras é pequeno e, comparativamente, possui maiores taxas de erros
	Pacific Biosciences	~1000	Ainda não está disponível	Não é esperado que seja liberado até o final de 2010/2011. Grande potencial (ver texto)

[a] Quando utilizado um *mate-pair run* que envolve a obtenção de sequências de ambas as extremidades das moléculas de DNA.

Como o sequenciador HeliScope detecta moléculas individuais, os moldes de sequenciamento podem ser densamente empacotados sobre a superfície do fluxo celular da Helicos (cerca de 100 milhões de moléculas moldes individuais por centímetro quadrado, ou bilhões a cada rodada). O sequenciamento ocorre por meio da reação de sequenciamento por síntese, o que envolve ciclos de adição de nucleotídeos individuais intercalados por passos de lavagem. Como resultado, as leituras da sequência individual são pequenas, mas a sequência total originada já é bastante respeitável (ver Tabela 8.1) e pode se tornar ainda maior no futuro, sendo que o custo dos reagentes é substancialmente menor do que para o sequenciamento que utiliza a reação de PCR para amplificar o DNA.

Outras tecnologias mais poderosas de sequenciamento de moléculas individuais estão sendo desenvolvidas e provavelmente promoverão uma transformação no campo do sequenciamento de DNA, tanto em termos de rapidez de sequenciamento como em custos reduzidos. Um dos pioneiros é um novo método revolucionário de sequenciamento de uma única molécula de DNA chamado SMRT (sequenciamento em tempo real de uma única molécula, do inglês ***single-molecule real-time sequencing***) que está sendo desenvolvido para sequenciar o DNA a uma velocidade 20 mil vezes maior do que os sequenciadores de segunda geração disponíveis atualmente no mercado. Ele também se baseia no sequenciamento por síntese, mas, ao contrário dos métodos descritos, a síntese ocorre em *tempo real*, para que as reações de sequenciamento sejam extremamente rápidas. Dentro das células, as DNA-polimerases realizam, naturalmente, a síntese de um novo DNA para duplicar genomas inteiros em minutos. O método SMRT rastreia visualmente DNA-polimerases individuais à medida que elas sintetizam as moléculas de DNA em tempo real. O sistema grava quais nucleotídeos são incorporados em cada posição utilizando nucleotídeos marcados com um dentre quatro fluoróforos diferentes, de acordo com a especificidade da base nitrogenada.

O SMRT explora duas inovações principais. Primeiro, utiliza dNTPs não convencionais marcados com fluoróforos. Na marcação normal do DNA, os fluoróforos se ligam às bases; à medida que cada nucleotídeo é inserido, o fluoróforo se torna uma parte permanente da fita de DNA. Uma vez que múltiplos nucleotídeos são inseridos no DNA, entretanto, o volume físico dos fluoróforos adicionados causa uma inibição estérica, inibindo a síntese posterior pela DNA-polimerase. Na maioria dos métodos que se baseiam no sequenciamento por síntese, portanto, o DNA é sintetizado um único nucleotídeo por vez, começando e parando a reação após cada incorporação. Grandes volumes de reagentes são requeridos e a *processividade* da polimerase, ou seja, a habilidade de continuar a sua função catalítica de forma repetida sem se dissociar do substrato, é fortemente limitada. No SMRT, em contrapartida, os fluoróforos são ligados ao grupo fosfato externo (γ) dentre os três resíduos de fosfato do nucleotídeo (**Figura 8.10**). Quando um dNTP ligado a um *fosfato marcado* pareia com a base complementar na fita molde, o sinal fluorescente pode ser gravado antes de a polimerase clivar o grupo trifosfato (**Figura 8.11B**). A clivagem do dNTP gera um dNMP não marcado que é incorporado na cadeia crescente de DNA para que a polimerase continue inserindo nucleotídeos sem problemas estéricos e possa gerar leituras com grandes comprimentos. Para maximizar o comprimento das leituras, é utilizada a DNA-polimerase ϕ29, a qual possui elevada processividade.

A segunda inovação chave é o *chip* SMRT, um filme de metal fino depositado sobre um substrato de vidro. O filme de metal contém milhares de pequenos orifícios com diâmetro de pequenos comprimentos de onda conhecidos como guias de onda em modo zero nanofotônicos (ZMWs; **Figura 8.11A**). Um ZMW individual possui um volume de apenas 20 zeptolitros (10^{-21} litros), e os ZMWs são constituídos por câmaras de visualização nanofotônicas, a partir das quais se pode observar diretamente uma DNA-polimerase através de um suporte de vidro, à medida que ela sequencia pela síntese um único molde de DNA. Nesse volume, a tecnologia pode detectar a atividade de uma única molécula brevemente imobilizada quando mantida no sítio ativo da enzima em um meio com milhares de nucleotídeos mar-

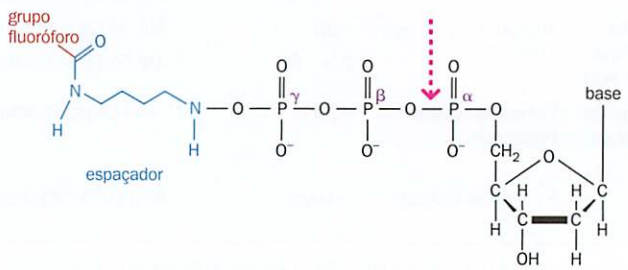

Figura 8.10 Um dNTP marcado com fluorescência ligado a um fosfato. A seta tracejada indica a clivagem que ocorre quando o nucleotídeo é incorporado ao DNA. Um grupo pirofosfato contendo os grupos β- e γ-fosfato e um fluoróforo ligado é destacado do resíduo dNMP que será incorporado ao DNA.

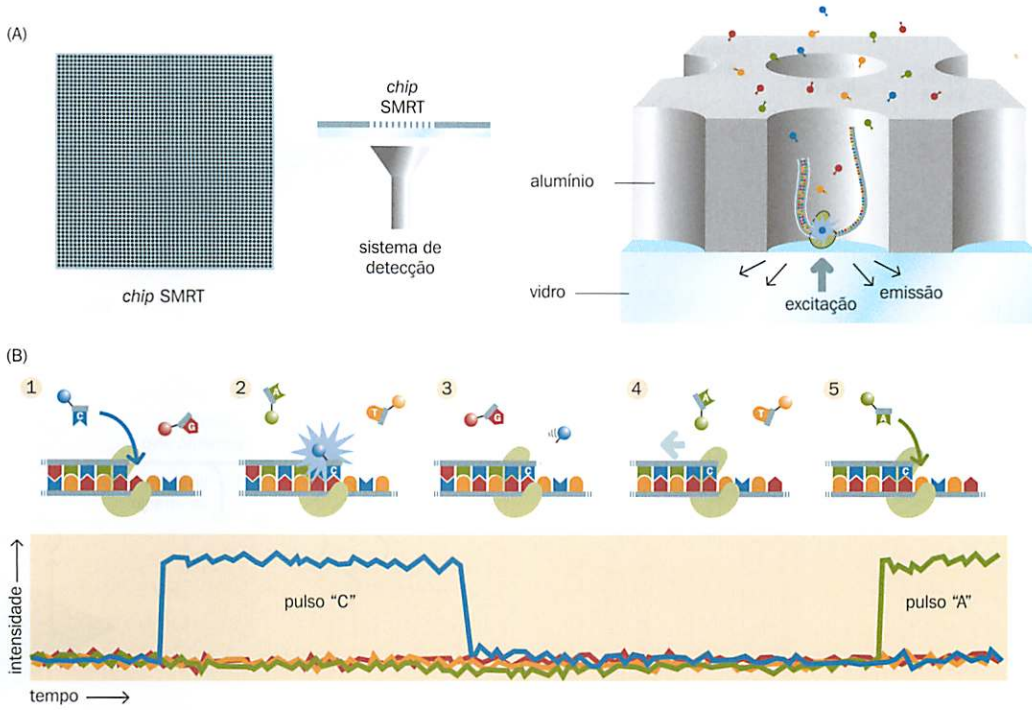

Figura 8.11 Princípios da tecnologia SMRT (sequenciamento em tempo real de uma única molécula, do inglês *single-molecule real-time sequencing*). (A) O *chip* SMRT e as câmaras de visualização nanofotônicas ZMW. O *chip* SMRT mostrado na esquerda é um filme de alumínio de 100 nm de espessura depositado sobre um substrato de silicone dioxidado que possui o tamanho de apenas 40 μm × 30 μm. O filme de metal contém milhares de perfurações conhecidas como guias de onda em modo zero (ZMWs) que podem agir com pequenas câmaras de visualização, em cada qual uma molécula de DNA-polimerase promove o sequenciamento de DNA com um único molde de DNA. O *chip* é montado próximo a um sistema de detecção de fluorescência. Com a visualização por meio do suporte de vidro pode-se gravar visualmente as reações de sequenciamento de DNA em tempo real em cada ZMW individual (direita). (B) Ciclos de incorporação de dNTP ligado a fosfato mais o intervalo de tempo esperado da intensidade fluorescente detectada. O SMRT é um método de sequenciamento por síntese que utiliza dNTPs marcadas com um entre quatro fluoróforos diferentes, dependendo de qual base está presente. Entretanto, como o fluoróforo está ancorado ao γ-fosfato dos dNTPs (ver Figura 8.10), ele não é incorporado à cadeia crescente de DNA. No entanto, uma molécula de DNA-polimerase irá manter um dNTP estável recém-chegado em seu sítio ativo por um breve período antes de o grupo trifosfato ser clivado e o dNMP ser inserido. Durante esse tempo, o fluoróforo anexado emite luz fluorescente, cuja cor corresponde à identidade da base. Os passos mostrados são os seguintes: (1) o nucleotídeo ligado ao fosfato começa a se associar com o molde no sítio ativo da polimerase, (2) causando uma elevação da fluorescência de saída no canal colorimétrico correspondente. (3) A formação da ponte fosfodiéster libera o produto do pirofosfato ligado ao marcador, o qual se difunde para fora do ZMW, terminando, assim, o pulso de fluorescência. (4) A polimerase transloca-se para a próxima posição e (5) o próximo nucleotídeo que será inserido se liga ao sítio ativo, dando início ao pulso subsequente. [Adaptada de Eid J, Fehr A, Gray J et al. (2009) *Science* 323, 133-138. Com permissão de American Association for the Advancement of Science.]

cados que se movem constantemente. Com refinamentos, a tecnologia é projetada para ser capaz de sequenciar o genoma humano em 1 hora por apenas US$ 100 dólares ou menos.

Outra grande área do desenvolvimento tecnológico no sequenciamento de moléculas únicas é o *sequenciamento com nanoporos*. Materiais como o silicone podem ser fabricados de tal maneira que contenham orifícios pequenos e muito finos (nanoporos). A princípio, nucleotídeos sucessivos em uma molécula individual fita simples de DNA são induzidos a passar através de um nanoporo. Enquanto passa através do poro, cada nucleotídeo bloqueia parcialmente o nanoporo de uma maneira diferente de acordo com sua característica, dependendo se for uma A, uma C, uma G ou uma T, e a quantidade de corrente elétrica que consegue passar através do nanoporo varia de acordo com o tipo de nucleotídeo que o está obstruindo. A mudança na corrente representa uma leitura direta da sequência de DNA. Está disponível um vídeo explicativo na página da internet em http://www.nanoporetech.com/sequences. Esta é uma área em rápido desenvolvimento, e assim como o SMRT, possui um grande potencial para oferecer, a baixo custo, sequenciamento de alto desempenho do DNA.

Métodos de captura de DNA por microarranjo permitem o ressequenciamento eficiente

Apesar do enorme progresso recente no sequenciamento do genoma, muitas aplicações atuais de sequenciamento massivo paralelo de DNA estão focadas em sequências alvo que, coletivamente, constituem uma pequena fração do genoma. Por exemplo, a triagem para genes de suscetibilidade ao câncer pode envolver o sequenciamento de todos os

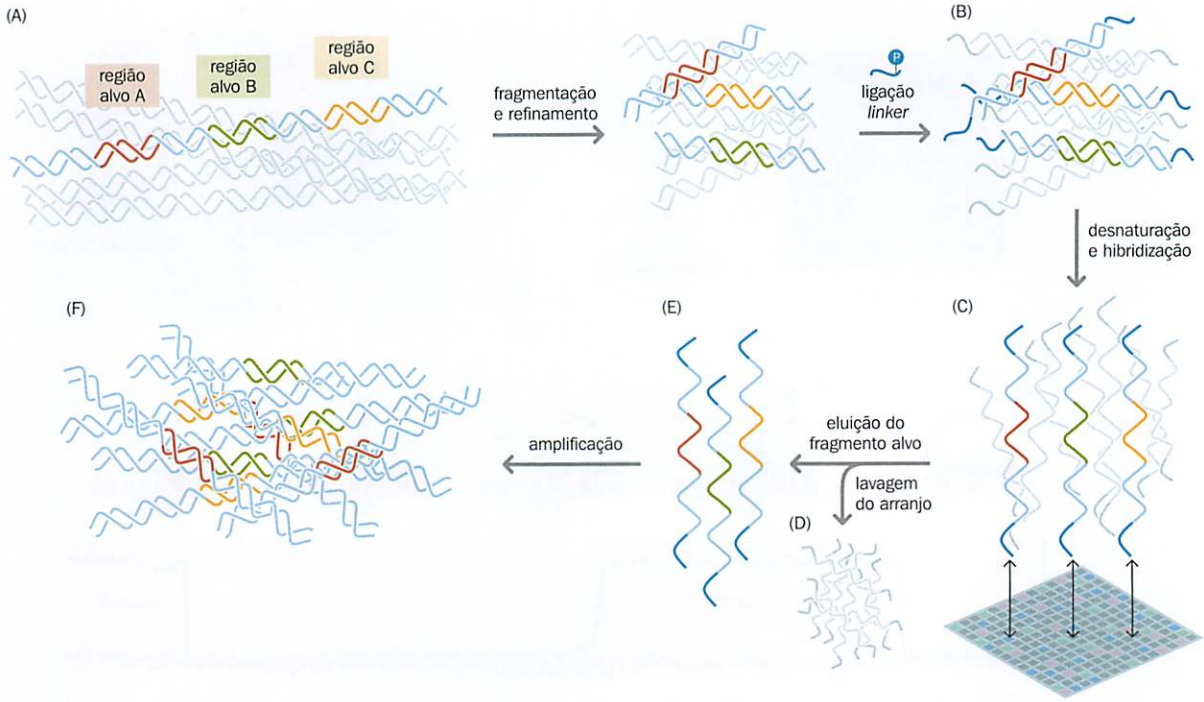

Figura 8.12 Captura de DNA baseada em arranjo com a tecnologia de captura de sequência NimbleGen$_{TM}$. O ressequenciamento de frações específicas do genoma pode ser feito, de modo conveniente, primeiramente pela captura de sequências-alvo desejadas por meio da hibridização seletiva. Para isso, microarranjos personalizados são desenvolvidos com sondas de DNA que representem apenas as regiões com a sequência-alvo desejada. (A) Um complexo de DNA inicial é fragmentado aleatoriamente e as extremidades são modificadas para gerar uma pequena população (cerca de 500 pb) de fragmentos com extremidades "cegas". (B) *Linkers* de oligonucleotídeos comuns são ligados às extremidades dos fragmentos. (C) O DNA é desnaturado e hibridizado a um microarranjo personalizado; após a hibridização, as sequências não ligadas são removidas pela lavagem (D). Os únicos fragmentos de DNA que estão ligados ao microarranjo devem ser as sequências-alvo desejadas de DNA, e elas podem ser recuperadas por eluição (E). As sequências-alvo recuperadas são, agora, amplificadas utilizando-se *primers* específicos para os *linkers* (F) para fornecer um *pool* amplificado enriquecido de sequências-alvo prontas para o sequenciamento massivo em paralelo. Para métodos alternativos de captura de DNA para sequenciamento massivo em paralelo, ver Mamanova et al. na seção de Leitura adicional. (Figura baseada, com permissão, em uma imagem fornecida pela Roche Inc.)

éxons, dos limites entre éxon-íntron e de elementos regulatórios conhecidos para todos os oncogenes conhecidos. Há centenas de oncogenes conhecidos, inclusive muitos genes novos identificados por programas internacionais como o *Cancer Genome Atlas*. A amplificação por PCR do que podem ser centenas de elementos de sequência em cada gene do câncer é tediosa e consome tempo. Como alternativa, é possível usar a hibridização por microarranjo como uma ferramenta para selecionar as sequências desejadas que são submetidas ao sequenciamento de alto desempenho.

O sistema de captura de sequências NimbleGen^{TM} é um sistema comercial que utiliza microarranjos de DNA para permitir o enriquecimento mediado por hibridização das sequências do DNA desejado dentro de um DNA inicial complexo, como o DNA genômico humano. Microarranjos são desenhados com oligonucleotídeos de todas as sequências-alvo desejadas e são hibridizados com o DNA inicial, randomicamente fragmentado, tratado de forma que oligonucleotídeos específicos (*linkers*) estejam ligados a ambas as extremidades e então desnaturados. Os fragmentos desejados dentro do DNA inicial devem hibridizar seletivamente o microarranjo e podem, posteriormente, ser recuperados por eluição, onde então *primers* específicos para os oligonucleotídeos *linkers* podem ser utilizados para amplificação do DNA por PCR na preparação para o sequenciamento massivo em paralelo (**Figura 8.12**). Agora é possível a captura do genoma em amplo espectro de, essencialmente, todos os éxons codificantes de proteínas, o **exoma** humano, para o ressequenciamento humano: o arranjo de exoma humano NimbleGen 2.1M pode capturar cerca de 180 mil éxons humanos e cerca de 550 sequências de miRNA (miRNA é um tipo de RNA funcional não codificante que regula a expressão de certos genes-alvo).

8.3 ANÁLISE DA ESTRUTURA DO GENOMA E PROJETOS GENOMA

Ao contrário dos genomas bacterianos, os quais consistem em um único tipo de uma pequena molécula comparativa de DNA, os genomas de organismos multicelulares complexos são compostos por várias e grandes moléculas de DNA. O genoma humano, por exemplo, é constituído por 25 diferentes tipos de moléculas de DNA. Há uma pequena molécula de DNA mitocondrial, a qual está presente em múltiplas cópias por célula e em 24 diferentes moléculas de DNA nucleares, com um tamanho médio de cerca de 135 Mb, que corresponde aos 24 diferentes cromossomos. Para começar a entender a estrutura e as funções do genoma humano, o ponto de partida necessário foi a obtenção da sua sequência. Até o início da década de 1980, entretanto, a ideia de sequenciamento do genoma humano parecia extremamente remota.

A sequência completa da molécula de DNA circular mitocondrial (mtDNA) humano foi, no entanto, logo publicada em abril de 1981, mas esta foi uma exceção. O mtDNA pode

ser facilmente purificado por ser encontrado unicamente no citoplasma e por ser comparativamente pequeno – algo em torno de 16,6 kb, o que corresponde a 1/200.000 do total do tamanho do genoma humano. Dados os esforços substanciais que foram necessários para sequenciar a pequena molécula de mtDNA, a tarefa de sequenciar moléculas de DNA cromossomais parecia irremediavelmente desencorajadora. Não apenas era sabido que as moléculas de DNA eram bastante grandes, como ainda não havia uma maneira fácil de separar um tipo de DNA cromossomal de outro.

O advento das bibliotecas de DNA ofereceu, a princípio, a possibilidade de sequenciamento de grandes genomas pela técnica de *shotgun*, onde clones selecionados aleatoriamente em uma biblioteca de DNA genômico são sequenciados até que o genoma inteiro seja coberto. Clones com insertos sobrepostos são normalmente gerados durante a construção das bibliotecas de DNA, fruto do resultado da fragmentação randômica do DNA. Cópias idênticas das mesmas moléculas de DNA cromossomais serão clivadas em diferentes sítios no DNA, assim qualquer pequena sequência única será representada por uma série de fragmentos de DNA sobrepostos de diferentes tamanhos produzidos pela clivagem diferencial do DNA.

Os diferentes fragmentos são, então, anexados a moléculas vetores idênticas e introduzidos aleatoriamente dentro de células (ver Figura 8.1). Se a média do tamanho do inserto de uma biblioteca genômica de DNA for selecionada para ser de algumas centenas de pares de bases, o sequenciamento padrão pelo método dideoxi pode ser utilizado para determinar os insertos em todos os clones. Se houver uma maneira acurada de identificar sequências com sobreposições, e clones suficientes forem sequenciados, torna-se possível sequenciar os genomas inteiros.

A abordagem descrita do sequenciamento inteiro do genoma pela técnica de *shotgun* (**Figura 8.13A**) é aplicada com mais sucesso para genomas pequenos; há grandes dificuldades em aplicar essa técnica para o sequenciamento inicial de grandes genomas de metazoários. O genoma humano, por exemplo, possui uma grande quantidade de sequências repetidas de DNA. Como os membros de uma família de DNA repetitivo podem ter a sequência bastante similar, é difícil mapear as sequências individuais.

Para contornar o problema do DNA repetitivo, uma estratégia diferente foi requerida para genomas de metazoários complexos. Foi necessário um tipo de arcabouço inicial, uma série de **mapas de *framework***, para ancorar sequências a regiões subcromossomais definidas como um prelúdio do sequenciamento por *shotgun* (**Figura 8.13B**). A questão crucial foi a seguinte: que tipo de mapa de *framework* poderia ser estabelecido para ajudar no sequenciamento do genoma humano?

Mapas de *framework* são necessários para o sequenciamento inicial de genomas complexos

Por décadas, geneticistas da área humana invejaram geneticistas que trabalhavam com organismos-modelo onde **mapas genéticos** clássicos de alta resolução puderam ser estabelecidos prontamente. Tais mapas foram baseados em genes mutantes: a partir do cruzamento de mutantes, a herança de fenótipos individuais pôde ser rastreada por gerações. Se dois fenótipos mutantes mostrassem uma tendência de serem herdados juntos, era esperado que os genes subjacentes estivessem razoavelmente próximos no mesmo cromossomo. A recombinação entre *loci* ligados poderia fornecer uma medida da distância física que separa os dois genes.

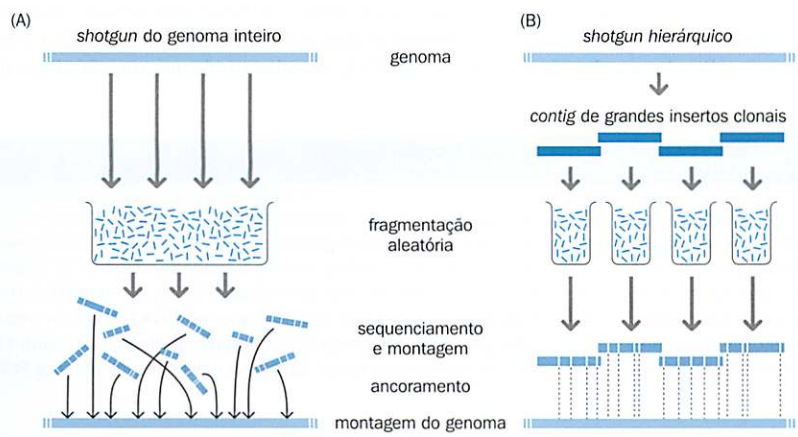

Figura 8.13 Duas estratégias para sequenciamento do genoma. (A) O sequenciamento por *shotgun* do genoma inteiro envolve a fragmentação indiscriminada do genoma em pequenos pedaços de DNA que são prontamente sequenciados. Isso gera rapidamente grandes quantidades de dados de sequência, mas o ancoramento das sequências em sítios específicos do genoma pode ser problemático. Grandes quantidades de DNA repetitivo em genomas complexos dificultam a localização, de modo não ambíguo, de regiões subcromossomais específicas. (B) Para o sequenciamento inicial de genomas complexos, é mais eficiente montar *contigs* de grandes insertos clonais para cada cromossomo e, então, fragmentar clones individuais em pedaços que são sequenciados para reconstruir a sequência do clone parental. [Adaptada de Waterston RH, Lander ES & Sulston JE (2002) *Proc. Natl Acad. Sci. USA* 99, 3712-3716. Com permissão de National Academy of Sciences, USA.]

Muito ocasionalmente, **mapas físicos** de alta resolução podem estar disponíveis. *Drosophila* foi o primeiro exemplo: o mapeamento genético não apenas foi simples, como mapas físicos de alta resolução puderam ser construídos utilizando-se cromossomos politênicos gigantes encontrados nas glândulas salivares de larvas de *Drosophila*. Os cromossomos politênicos são extremamente fora do comum, pois podem ser visualizados sob o microscópio durante a interfase, quando os cromossomos estão altamente estendidos. Isso é possível devido ao fato de que eles sofreram numerosos passos de replicação do DNA sem que ocorresse divisão celular (endomitose). Como resultado, cada cromossomo é constituído por 1.000 ou mais cromátides que se mantêm unidas umas às outras ao longo do seu comprimento, como canudos fortemente empacotados em uma caixa.

Por razões éticas e práticas, o mapeamento genético clássico nunca poderia ser contemplado em humanos. O único mapa físico que está disponível foi baseado em cromossomos distintos de acordo com o tamanho e a forma, utilizando certas linhagens para produzir padrões de bandeamento subcromossomal (ver Figura 2.15). Clones de bibliotecas de DNA poderiam ser mapeados em regiões subcromossomais, por exemplo, por hibridização ao DNA cromossomal desnaturado (hibridização *in situ*), mas o processo de mapeamento seria lento, e a resolução subcromossomal não seria particularmente elevada.

O avanço que pavimentou a maneira de mapear o genoma humano surgiu com a percepção de que mapas genéticos não precisam ser baseados em genes. A mutação é um processo essencialmente randômico. Para qualquer tipo de cromossomo humano, a sequência de DNA varia de uma cópia cromossomal para outra em cerca de um nucleotídeo dentre milhares, em média. Apenas uma pequena fração dessa variação causa fenótipos mutantes, e o grande volume das mudanças é encontrado em regiões entre genes ou dentro de íntrons. Uma vez que os ensaios experimentais tenham sido desenvolvidos para rastrear esse tipo geral de *polimorfismo de DNA*, mapas de *framework* baseados em marcadores de DNA puderam ser estabelecidos.

Após a criação dos mapas de *framework* iniciais baseados em marcadores genéticos, mapas marcadores de DNA de alta densidade foram desenvolvidos, nos quais muitos dos marcadores não eram polimórficos, mas foram simplesmente escolhidos por terem uma sequência única que poderia ser utilizada no ensaio do PCR. Uma vez que *mapas marcadores de framework* com uma adequada densidade foram desenvolvidos, foi possível construir *mapas clonais de framework*. Como descrito na próxima seção, mapas clonais de DNA envolvem a identificação e o arranjo de clones de DNA em uma ordem linear que corresponde à ordem linear dentro dos cromossomos das sequências de DNA clonadas. Mapas clonais abrangentes para cada cromossomo forneceram o substrato final para o sequenciamento do genoma que remete, em última análise, ao mapa físico a uma resolução de 1 pb.

A ordem linear dos clones de DNA genômico em um *contig* está de acordo com as suas localizações subcromossomais originais

Os substratos perfeitos para sequenciar genomas complexos são grupos de clones de DNA genômico que contêm grandes insertos ordenados de forma linear, de acordo com a origem subcromossomal dos seus insertos. Uma série de clones em que o inserto de cada um se sobrepõe parcialmente aos insertos vizinhos é conhecida como ***contig* clonal**: representa uma sequência de DNA contígua (contínua) de uma região subcromossomal ou de um cromossomo inteiro (**Quadro 8.1**).

Clones com insertos sobrepostos são gerados, normalmente, durante a construção das bibliotecas de DNA genômico, pois primeiramente o DNA é submetido à fragmentação *randômica* (ver Figura 8.1). Cópias idênticas das mesmas moléculas de DNA cromossomal serão clivadas em locais diferentes no DNA, e assim qualquer sequência pequena

QUADRO 8.1 Mapas de *framework* baseados em marcadores de DNA e *contigs* clonais

O primeiro sequenciamento de genomas complexos (os quais possuem, invariavelmente, grandes quantidades de DNA repetitivo) é auxiliado, primeiramente, por meio da construção de **mapas de *framework*** para cada molécula de DNA cromossomal. Os mapas de *framework* mais adequados para o sequenciamento do DNA são baseados em ***contigs* clonais**, séries de fragmentos de DNA clonados, dispostos linearmente, que sobrepostos representam, coletivamente, sequências de DNA cromossomal (**Figura 1**).

A montagem de *contigs* clonais depende da capacidade de se identificar sobreposições entre fragmentos de DNA clonados. Há várias maneiras de se fazer isso, porém, a mais frequente é verificar a presença de certos marcadores de DNA conhecidos, a fim de mapear a região subcromossomal aproximada. Um marcador de DNA é qualquer sequência de DNA pequena que possa ser analisada de alguma maneira, tanto por ensaio de hibridização como, de forma mais conveniente, por PCR.

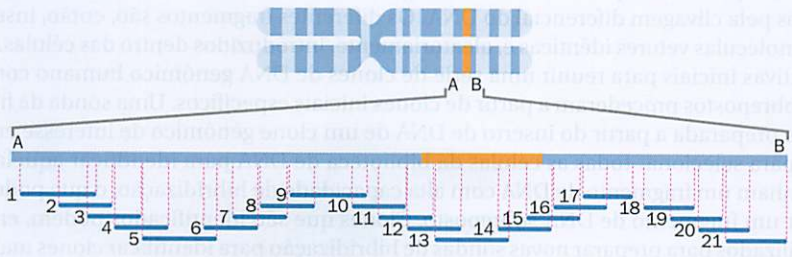

Figura 1 Um *contig* clonal específico. A sequência de DNA cromossomal da posição A à posição B é representada pelo sobreposição de insertos de DNA em uma série linear de clones de DNA genômico. Clones com insertos sobrepostos são gerados por meio da fragmentação aleatória do DNA quando uma biblioteca de DNA genômico é construída (ver Figura 8.1).

Marcadores de DNA podem, ou não, ser polimórficos (Tabela 1), mas para serem úteis para propostas de mapeamento, um marcador precisa ter uma localização cromossomal singular. Diferentes métodos têm sido utilizados para mapear um marcador de DNA humano em uma localização subcromossomal específica. Uma abordagem comum é a montagem de painéis de **células híbridas** construídas artificialmente que contenham um grupo total de cromossomos de roedores mais um, ou mais, cromossomos humanos específicos ou múltiplos fragmentos de cromossomos humanos. Estes últimos são gerados por meio da exposição de células humanas à radiação controlada, causando quebra cromossomal e, então, fusionando as células humanas com as células de roedores, resultando em grupos variáveis de fragmentos cromossomais humanos que possam inserir-se dentro dos cromossomos de roedores (*radiação híbrida*). De forma alternativa, marcadores de DNA têm sido mapeados em cromossomos, por meio da marcação com fluoróforo de clones conhecidos por conter o marcador, sendo hibridizados às preparações cromossômicas humanas fixadas e tratadas adequadamente sobre lâminas de microscópio (FISH de cromossomos; ver Figuras 2.16A e 2.17A).

Mapas de marcadores de DNA podem ser construídos de diferentes maneiras. Marcadores polimórficos podem, simplesmente, ser analisados em múltiplas gerações de indivíduos de uma linhagem a fim de identificar grupos de marcadores ligados que devem estar no mesmo cromossomo. Marcadores não polimórficos podem ser dispostos em mapas de marcadores utilizando o método de PCR para analisá-los em painéis de células híbridas irradiadas, ou em painéis de clones YAC ou outros clones de DNA genômico com grandes insertos. Marcadores comuns são **sítios marcadores de sequências (STS)** originados de clones que tenham sido, no mínimo, parcialmente sequenciados, inclusive clones com o DNA sequenciado e identificados previamente, como clones gênicos ou outros clones de DNA de interesse que já tenham sido bem caracterizados e mapeados. De modo mais geral, é possível sequenciar as extremidades de vários clones com grandes insertos a partir da biblioteca de DNA genômico, gerando assim milhares de marcadores.

Se um grande número de marcadores STS está disponível, eles podem ser utilizados para construir mapas marcadores de *framework* que permitem que os clones sejam unidos com mapas de *contigs* clonais. Um exemplo antigo de um pequeno mapa de *contig* clonal que foi montado a partir do mapeamento de marcadores STS é mostrado na **Figura 2**. A ideia é agrupar os clones em uma ordem linear de acordo com a presença ou ausência de marcadores STS. Clones de cromossomos artificiais de leveduras eram, inicialmente, uma escolha popular para montagem de mapas de *contigs* clonais, mas YACs contendo grandes insertos humanos se mostraram instáveis, sendo suscetíveis a rearranjos internos e deleções (ver Figura 2).

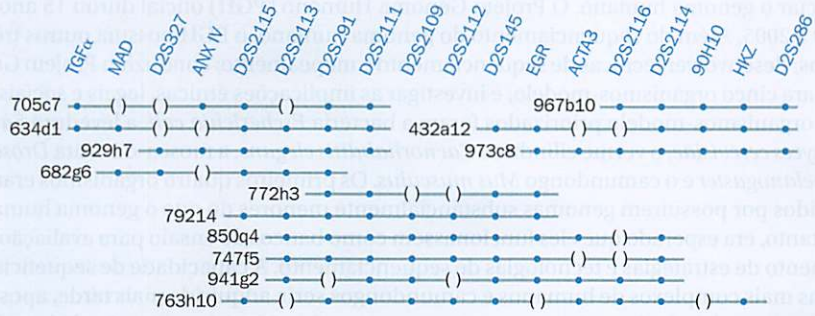

Figura 2 Um exemplo prático de um *contig* clonal humano precoce montado por meio do mapeamento por STS (sítios marcadores de sequências). Este exemplo demonstra um *contig* clonal de YAC representando uma porção do cromossomo 2 humano. Os clones são mostrados como linhas horizontais com os seus respectivos nomes à esquerda. Marcadores STS, na parte superior, incluem alguns marcadores derivados de genes (TGFα, MAD, ANX IV, etc.) e alguns marcadores anônimos do cromossomo 2 (D2S327, D2S2115, etc.). A tipagem positiva para um marcador STS é indicada por um círculo fechado; parênteses indicam a ausência de um STS esperado (provavelmente, devido à instabilidade do YAC).

Tabela 1 Marcadores comuns utilizados na construção de mapas de DNA de framework de genomas complexos

Tipo de marcador	Marcador	Definição
Polimórfico	Polimorfismo do tamanho do fragmento de restrição (RFLP)	Qualquer polimorfismo de DNA que resulte na criação ou destruição de uma sequência reconhecida por uma endonuclease de restrição específica. Utilizado para análises por hibridização, mas agora também é utilizado no ensaio por PCR
	Microssatélite	Uma sequência contendo várias (normalmente 10 ou mais) repetições em *tandem* de uma sequência de 1-4 nucleotídeos. Encontrada frequentemente em genomas complexos e altamente polimórficos
	Polimorfismo de nucleotídeo único (SNP)	Uma mudança em um único nucleotídeo é polimórfica. O polimorfismo é normalmente limitado (apenas dois alelos são significativamente frequentes), mas SNPs são bastante adequados para utilização na genotipagem automatizada, como no uso do pirossequenciamento (ver Figura 8.8)
Não polimórfico	Sítios marcadores de sequências (STS)	Qualquer sequência de DNA que tenha sido mapeada em uma localização subcromossomal e que possa ser amplificada por PCR
	Etiqueta de sequência expressa (EST)	Um subgrupo de sítios marcadores de sequência que se localizam dentro de sequências de DNA conhecidas por serem transcritas

e única será representada por uma série de fragmentos de DNA de diferentes tamanhos produzidos pela clivagem diferencial do DNA. Os diferentes fragmentos são, então, inseridos em moléculas vetores idênticas e, aleatoriamente, introduzidos dentro das células.

Tentativas iniciais para reunir uma série de clones de DNA genômico humano com insertos sobrepostos procederam a partir de clones iniciais específicos. Uma sonda de hibridização preparada a partir do inserto de DNA de um clone genômico de interesse era utilizado para selecionar todas as células da biblioteca de DNA para identificar aquelas que continham um fragmento de DNA com alta capacidade de hibridização, o que poderia indicar um fragmento de DNA sobreposto. Clones que são identificados podem, então, ser utilizados para preparar novas sondas de hibridização para identificar clones mais sobrepostos que estão mais distantes. Esse procedimento, conhecido como *caminhada cromossômica* (*chromosome walking*), era extremamente demorado.

De modo mais eficiente, métodos de identificação de clones que cobrem o genoma inteiro utilizando insertos sobrepostos foram desenvolvidos, subsequentemente, utilizando o ***fingerprint* clonal**. A ideia era submeter *todos* os clones da biblioteca a algum ensaio que pudesse ajudar a identificar clones com insertos de DNA sobrepostos. Por exemplo, insertos de DNAs clonais foram submetidos ao mapeamento por restrição e análise do conteúdo de DNA repetitivo, em busca de padrões espaciais característicos de sítios para endonucleases de restrição ou de sequências repetitivas particulares.

Impressões digitais (*fingerprinting*) de clones baseadas no mapeamento de restrição e/ou de conteúdo de DNA repetitivo ainda são bastante trabalhosas. Um método ainda mais eficiente depende da disponibilidade de um mapa de alta densidade de marcadores de DNA. Vários marcadores de DNA podem ser utilizados. Eles podem ser polimórficos (possivelmente localizados sobre um mapa genético) ou não polimórficos. Porém há dois requisitos: devem possuir uma localização subcromossomal única e serem passíveis de algum tipo de análise (ver Quadro 8.1).

O Projeto Genoma Humano foi um empenho internacional e o primeiro Grande Projeto biológico

A percepção no início da década de 1980 de que mesmo grandes genomas, como o genoma humano, poderiam ser sequenciados deflagrou rigorosos esforços no planejamento para sequenciar o genoma humano. O Projeto Genoma Humano (PGH) oficial durou 15 anos, de 1990 a 2005. Além do sequenciamento do genoma humano, o PGH possuía outros três objetivos: desenvolver técnicas de sequenciamento e mapeamento; conduzir o Projeto Genoma para cinco organismos-modelo; e investigar as implicações étnicas, legais e sociais.

Os organismos-modelo priorizados foram a bactéria *Escherichia coli*, a levedura *Saccharomyces cerevisiae*, o verme cilíndrico *Caenorhabditis elegans*, a mosca-da-fruta *Drosophila melanogaster* e o camundongo *Mus musculus*. Os primeiros quatro organismos eram conhecidos por possuírem genomas substancialmente menores do que o genoma humano; portanto, era esperado que eles funcionassem como bancos de ensaio para avaliação e refinamento de estratégias e tecnologias de sequenciamento. A capacidade de sequenciar genomas mais complexos de humanos e camundongos seria adquirida mais tarde, após a aprendizagem advinda de projetos genoma menores assim como da utilização do benefício máximo do aperfeiçoamento tecnológico.

Devido à grande escala envolvida, o PGH foi o primeiro Grande Projeto da biologia. O objetivo final era obter uma tabela periódica para biologia, baseada em genes em vez de elementos químicos. O financiamento público do PGH também requereu um verdadeiro esforço internacional. Ele veio a ser representado pelo Consórcio Internacional de Sequenciamento do Genoma Humano de 20 diferentes centros nos EUA, Reino Unido, França, Japão, Alemanha e China e foi marcado pelo amplo compartilhamento de dados e pela colaboração na estratégia, metodologia e análise dos dados.

Muito da tecnologia de sequenciamento do genoma foi concentrado em alguns poucos centros de sequenciamento e mapeamento de grandes genomas, os quais possuíam recursos em escala industrial e uma grande capacidade de análise dos dados. Juntamente a esses centros, uma rede de pequenos laboratórios ao redor do mundo inteiro interagia na tentativa principal de mapear e identificar genes ligados a doenças e, geralmente, focando-se em regiões subcromossomais bastante específicas. Além dos esforços no financiamento público das pesquisas, programas de financiamento privado possuíam, em paralelo, projetos de sequenciamento similares.

Os primeiros mapas de *framework* do genoma humano eram mapas genéticos de cromossomos individuais. Os mapas genéticos eram de baixa resolução, mas forneceram o

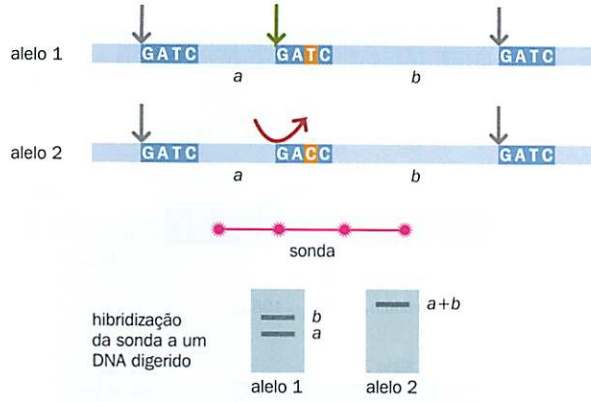

Figura 8.14 Polimorfismo do tamanho do fragmento de restrição (RFLP). Um RFLP é qualquer polimorfismo de DNA que causa uma mudança na sequência de reconhecimento de uma endonuclease de restrição. Frequentemente, isso é o resultado de uma mudança nucleotídica simples, como a transição T → C, aqui demonstrada, a qual permite que o alelo 1 tenha uma sequência de reconhecimento (GATC) para *Mbo*I que não é encontrada no alelo 2. A sonda de hibridização aqui ilustrada detectaria as sequências *a* e b, sendo que, se o DNA genômico é digerido com *Mbo*I antes da hibridização, a sonda pode distinguir o alelo 1 (o corte produz dois fragmentos *a* e *b*) e o alelo 2 (*a* e *b* estão no mesmo fragmento). De forma alternativa, mas não mostrada aqui, os alelos podem ser convenientemente analisados por PCR utilizando-se um *primer* a montante derivado da sequência *a* e a jusante um *primer* derivado da sequência *b*. Os produtos da amplificação são digeridos com *Mbo*I e fracionados de acordo com o tamanho para distinguir um alelo do outro.

esqueleto sobre o qual uma série de mapas físicos mais detalhados foram construídos, culminando em *contigs* clonais para cada cromossomo, os quais foram então utilizados para o estágio de sequenciamento final.

Os primeiros mapas genéticos eram de baixa resolução e foram construídos, em sua maioria, com marcadores de DNA anônimos

No mapeamento genético, diferentes marcadores polimórficos são analisados em todos os membros de uma variedade de famílias com múltiplas gerações. Os genótipos resultantes são então analisados computacionalmente para descobrir a maneira como os alelos de diferentes marcadores segregam em cada evento meiótico, conectando parentes à sua descendência, e para identificar marcadores onde alelos específicos cossegregam. A descoberta de, aparentemente, polimorfismos de DNA randômicos no genoma humano suscitou a ideia de construir um mapa genético humano mais amplo e não clássico, que fosse baseado predominantemente em marcadores de DNA anônimos em vez de genes.

O primeiro mapa de ligação genético do genoma humano foi publicado em 1987 e era baseado em **polimorfismos do tamanho do fragmento de restrição (RFLPs)**. Um RFLP é qualquer polimorfismo de DNA, frequentemente uma mudança de um único nucleotídeo, que resulta na criação ou na destruição de uma sequência de reconhecimento para uma endonuclease de restrição específica (**Figura 8.14**). O RFLP pode ser analisado por meio da digestão de DNA com a endonuclease de restrição assim como pela utilização de uma sonda de hibridização que cubra a sequência variável (ou imediatamente adjacente a ela) ou, mais convenientemente, *primers* flanqueadores da sequência variável para amplificar a sequência requerida e, então, digeri-la com endonuclease de restrição.

Embora tenha sido uma grande conquista, o primeiro mapa genético tinha seu uso limitado. Havia muito poucos marcadores – somente 393 RFLPs foram utilizados. Além disso, os marcadores não eram muito polimórficos, pois RFLPs possuem apenas dois alelos – ou o sítio está presente ou não. Para criar um mapa genético útil, um mapa de alta densidade foi necessário, com marcadores que fossem mais polimórficos.

Mapas genéticos de segunda geração eram baseados em diferentes classes de marcador de DNA. **DNA microssatélite** é o nome dado às pequenas sequências repetidas em *tandem*, com uma unidade de repetição de um a quatro nucleotídeos. Eles são comuns no genoma humano. Microssatélites individuais muitas vezes possuem um grande número de repetições, tornando-os instáveis, variando o número de repetições.

A instabilidade do microssatélite surge porque durante a replicação do DNA, a DNA-polimerase frequentemente comete erros ao copiar as longas repetições em *tandem*. Algumas vezes a polimerase deixa para trás uma unidade de repetição individual e falha em copiá-la; ela também pode voltar atrás por engano e copiar a mesma unidade de repetição individual duas vezes. Como resultado, microssatélites são altamente polimórficos, e os múltiplos alelos podem ser distinguidos utilizando-se o ensaio por PCR (**Figura 8.15**). O primeiro mapa de ligação humano baseado em marcadores de microssatélite foi publicado em 1992 utilizando-se 814 marcadores (corresponde a aproximadamente um marcador a cada 3,5 Mb). Já em 1994, um mapa genético integrado (baseado principalmente em microssatélites, mas contendo alguns outros marcadores) possuía uma densidade de aproximadamente um marcador por megabase.

O mapa de 1994 teve uma resolução suficientemente alta para alcançar os objetivos do mapeamento genético do PGH e, daí em diante, o principal foco foi no desenvolvimento

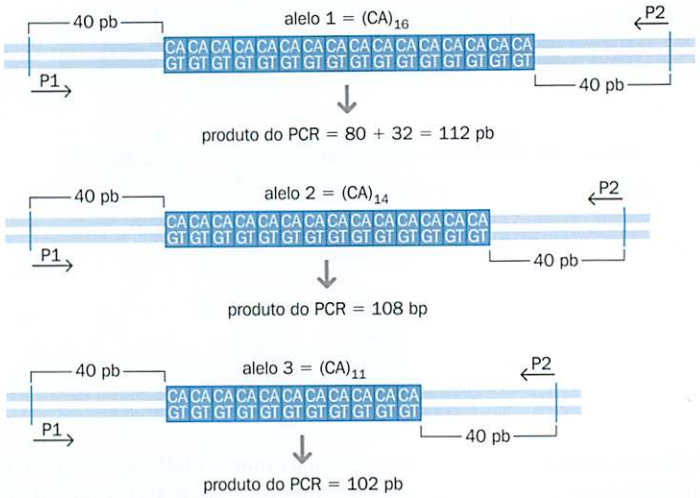

Figura 8.15 Desenvolvendo um ensaio de PCR para polimorfismos de microssatélites de DNA. Um marcador de microssatélite de DNA é um tipo de polimorfismo de DNA no qual variações no número de pequenas repetições em *tandem* (neste caso, uma repetição dinucleotídica de [CA]/[TG]) produzem polimorfismos de tamanho. Neste exemplo, a extremidade 5′ do *primer* P1 a montante e do *primer* P2 a jusante estão localizadas a 40 pb do arranjo de microssatélite e, portanto, o tamanho do fragmento amplificado será a soma de 80 + o tamanho do arranjo (32, 28 ou 22 pb neste exemplo). Os diferentes alelos podem ser separados por eletroforese em gel de poliacrilamida.

e refinamento de mapas físicos levando à produção de um último mapa físico, o da sequência completa de cada cromossomo. Entretanto, o mapeamento genético humano continuou de duas maneiras principais. Primeiro, mapas de microssatélites de resoluções ainda mais superiores foram desenvolvidos. Segundo, mapas de alta densidade foram desenvolvidos para polimorfismos de nucleotídeos únicos (SNPs) pelo Consórcio Internacional HapMap (mapeamento de haplótipo). A principal motivação tem sido a identificação de genes comuns a doenças. Os SNPs não são particularmente polimórficos (normalmente possuem dois alelos), mas são bastante adequados para tipagem automática, por métodos como pirossequenciamento (ver Figura 8.8). O mapa de SNP mais recente tem cerca de 3 milhões de marcadores de SNP, ou aproximadamente um SNP por quilobase.

Mapas físicos do genoma humano progrediram a partir de mapas de marcadores para mapas de *contigs* clonais

Diversos tipos de mapas físicos foram construídos para o genoma humano (**Tabela 8.2**), mas um objetivo principal do PGH foi construir um mapa de *contigs* clonais para cada cromossomo, como um prelúdio para o sequenciamento do genoma. Para isso, bibliotecas genômicas contendo grandes insertos de DNA foram escolhidas e, inicialmente, bibliotecas de cromossomos artificiais de levedura (YACs) foram selecionadas por conseguirem acomodar grandes insertos.

Em 1993, Daniel Cohen e colegas do laboratório do *Centre d'Études du Polymorphisme Humaine* (CEPH), em Paris, reportaram uma biblioteca de YAC humana com mais de

TABELA 8.2 Diferentes tipos de mapas físicos utilizados para mapear o genoma nuclear humano

Tipo de mapa	Exemplos/metodologia	Resolução
Citogenético	mapas de bandeamento de cromossomos	uma banda média possui vários Mb de DNA
Mapas de cromossomos com ponto de quebra	painéis híbridos de células somáticas com fragmentos cromossômicos originados a partir de cromossomos com translocações, ou deleções, naturais	a distância entre pontos de quebra cromossomais adjacentes em um cromossomo é de, geralmente, várias Mb
	mapas híbridos de radiação (HR) monocromossomal	a distância entre os pontos de quebra é de, frequentemente, muitas Mb
	mapas HR do genoma inteiro	a resolução pode ser de até 0,5 Mb
Mapas de restrição	criados com endonucleases de restrição que possuem baixa frequência de clivagem	várias centenas de kb
Mapas de *contigs* clonais	clones YAC com sobreposição	a média de um inserto YAC é de várias centenas de kb
	clones BAC com sobreposição	a média de um inserto BAC é de 160 kb
Mapas STS	requer a informação prévia da sequência a partir dos clones que tenham sido mapeados em regiões subcromossomais	menos de 1 kb é possível, mas mapas STS padrão possuem resolução de dezenas de kb
Mapas EST	requer o sequenciamento do cDNA e, então, o mapeamento de cDNAs de volta em outros mapas físicos	resolução média no genoma nuclear humano é de ~90 kb
Mapas de sequência de DNA	sequência nucleotídica completa do DNA de cada cromossomo	1 pb

STS, sítio marcador de sequência; EST, etiqueta de sequência expressa.

30 mil clones independentes e uma média de tamanho de inserto de pouco menos de 1 Mb. Estabelecendo sobreposição de clones a partir de *fingerprinting* clonal eles foram, por fim, capazes de reunir um mapa de *contigs* de YAC que cobria cerca de 75% do genoma, com uma média de tamanho de *contig* de cerca de 10 Mb.

Outro importante objetivo do mapeamento físico foi construir mapas baseados em **sítio marcador de sequência (STS)**. Um marcador de STS é qualquer sequência única de DNA conhecida que pode ser facilmente analisada por PCR. Os marcadores de STS incluem marcadores polimórficos, como microssatélites que podem ser facilmente genotipados por PCR e um número potencialmente grande de marcadores não polimórficos. O sequenciamento genômico randômico de clones de DNA forneceu muitos marcadores de STS não polimórficos; outros foram obtidos a partir de genes sequenciados e clones de cDNA, os quais foram denominados **etiquetas de sequência expressa (EST)**.

Em 1995, uma colaboração entre o Instituto de Tecnologia de Massachusetts (MIT) e o Instituto Whitehead tornou público uma mapa humano de STS contendo mais de 15 mil marcadores STS com um espaçamento médio de menos de 200 kb. Neste marco, as localizações cromossomais de marcadores STS não polimórficos foram obtidas por duas abordagens. A primeira abordagem utilizou o mapeamento de conteúdo de STS. Clones YAC da biblioteca CEPH YAC foram analisados para um enorme número de marcadores STS a fim de identificar clones que compartilhassem certos marcadores, para atribuir-lhes *contigs* de YAC com uma localização subcromossomal conhecida. A segunda abordagem utilizou mapeamento de células híbridas. Diferentes marcadores de STS foram analisados em painéis de **células híbridas** de humanos/roedores. Cada célula híbrida continha um complemento total de cromossomos de roedores, mas diferia com relação ao conteúdo total de cromossomos humanos ou diferentes grupos de fragmentos de cromossomos humanos que tinham se integrado aos cromossomos de roedores.

Embora o sucesso do mapeamento de YAC tenha sido impressionante, havia um problema. Os YACs contendo grandes insertos de DNA humano (os quais irão conter muita sequência de DNA repetitivo) se mostraram suscetíveis a rearranjos e deleções (ver Figura 2 no Quadro 8.1). Como os insertos YAC não eram, frequentemente, representações fidedignas do DNA humano inicial original, clones YAC não puderam servir como modelo para os esforços finais do sequenciamento do genoma.

Sistemas alternativos com grandes insertos clonais foram desenvolvidos. Bibliotecas de cromossomo artificial de bactéria (BAC) e cromossomo de fago P1 artificial (PAC) possuem insertos menores (80-250 kb) do que YACs, mas, mais importante, insertos humanos são comparativamente estáveis em BACs e PACs. Para fornecer ainda mais marcadores de STS, clones de bibliotecas BAC foram submetidos rotineiramente ao sequenciamento final, em que algumas centenas de nucleotídeos eram sequenciados até o final do inserto. Por fim, grandes *contigs* clonais BAC/PAC foram estabelecidos para cada cromossomo humano, pavimentando o caminho para a fase final do sequenciamento do genoma (ver visão geral na **Figura 8.16**).

A fase final de sequenciamento do Projeto Genoma Humano foi uma corrida para um término precoce

O progresso do Projeto Genoma Humano foi mais rápido do que o esperado (ver Quadro 8.2 para a linha do tempo). Mapas genéticos foram desenvolvidos antes do tempo estimado, e o estágio final do sequenciamento de DNA em larga escala foi facilitado por desenvolvimentos no sequenciamento do DNA baseado em fluorescência. A competição entre programas de fomento públicos e privados também direcionou para uma rápida fase final.

Uma importante lógica para o PGH – e uma grande motivação para o financiamento privado de sequenciamento do genoma – foi a capacidade de estudar genes humanos. Começando no início de 1990, foram feitas tentativas para a obtenção de sequências parciais das regiões 3′ não traduzidas do maior número de clones de cDNA possível, gerando um grande número de ESTs. Íntrons são raramente encontrados em regiões 3′ não traduzidas de genes humanos e, portanto, um ensaio por PCR baseado na sequência EST pode ser utilizado para escrever o DNA genômico.

Subsequentemente, o mapeamento em larga escala de ESTs contra painéis híbridos de radiação produziram o primeiro mapa abrangente de genes humanos. O mapa gênico resultante foi publicado em 1998 (ver Quadro 8.2) e pareceu identificar as posições de 30 mil genes humanos. Entretanto, a extensão total dos genes não pôde ser conhecida com maior precisão até que a sequência do genoma fosse entregue.

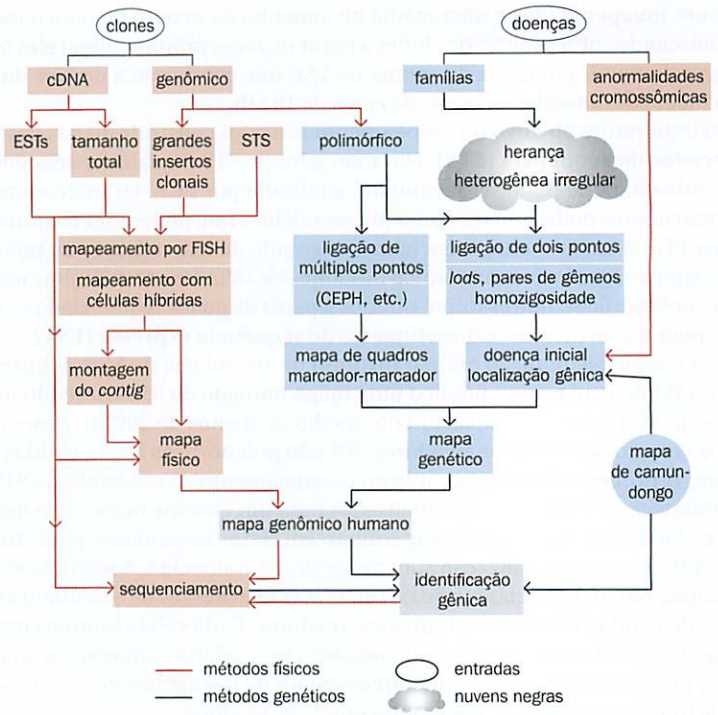

Figura 8.16 Principais estratégias científicas e abordagens utilizadas no Projeto Genoma Humano (PGH). O PGH exigiu o isolamento de clones genômicos humanos de cDNA. Os clones foram utilizados para construir mapas genéticos e físicos de alta resolução que pavimentaram o caminho para o sequenciamento do genoma. Inevitavelmente, o PGH interagiu com pesquisas referentes ao mapeamento e à identificação de genes associados a doenças em humanos. Os dados produzidos foram canalizados em bancos de dados de mapeamento e sequências, permitindo o rápido acesso eletrônico e análises de dados. Projetos auxiliares (não mostrados aqui) incluíram o estudo da variação genética, projetos genoma para organismos-modelo e pesquisas sobre implicações éticas, legais e sociais. EST, etiqueta de sequência expressa; CEPH, *Centre d'Études du Polymorphisme Humaine*; *lods*, *lod score* (logaritmo da probabilidade); STS, sítios marcadores de sequências; FISH, hibridização *in situ* com fluorescência.

Já no estágio inicial estava claro que os genes humanos não estavam distribuídos uniformemente ao longo, ou entre, cromossomos. Alguns cromossomos eram ricos em genes; outros eram pobres (**Figura 8.17**). As regiões heterocromáticas do genoma – incluindo a maior parte do cromossomo Y e regiões substanciais dos cromossomos 1, 9, 16 e assim por diante – eram conhecidas por serem essencialmente desprovidas de genes e extraordinariamente ricas em DNA repetitivo, o que tornaria o mapeamento extremamente difícil. Como resultado, o PGH se focou quase que exclusivamente nas regiões eucromáticas restantes que, coletivamente, correspondiam a cerca de 90% do genoma humano.

Para o PGH, que era amparado pelo financiamento público, a maioria das sequências foi concedida por grandes centros genômicos, em especial pelo Instituto Wellcome Trust Sanger, Universidade de Washington, Baylor College of Medicine, MIT/Harvard, Instituto de Genoma DoE Joint, Instituto Japonês RIKEN e pelo centro French Genoscope. Para assegurar a eficiência, ficou acordado que centros específicos teriam a responsabilidade primária pela montagem de *contigs* clonais e subsequente sequenciamento de cromossomos individuais: por exemplo, o Instituto Sanger para o cromossomo 1 e a Universidade de Washington para o cromossomo 2.

O PGH amparado pelo financiamento público encontrou forte competição do sequenciamento do genoma privado. Em 1999, a companhia Celera anunciou que pretendia produzir um esboço da sequência do genoma humano em 2 anos e que o faria a partir do sequenciamento por *shotgun* de todo o genoma (ver Figura 8.13A), em vez da abordagem mais lenta de montagem dos *contigs* utilizada pelo PGH.

A corrida subsequente entre o Consórcio Internacional para Sequenciamento Genoma Humano (IHGSC), financiado publicamente, e o financiamento privado Celera acelerou o PGH. Em 2001, ambos os lados publicaram um esboço da sequência do genoma humano que cobria cerca de 90% da sequência do genoma eucariótico. O componente eucromático é cerca de 90% do genoma total, portanto, o esboço das sequências é, na verdade, uma representação de cerca de 80% do genoma total, mas 90% da sequência gênica total.

Embora a corrida tenha parecido terminar empatada em 2001, não foi uma corrida justa, e a linha de chegada não havia sido alcançada. O IHGSC tornou seus dados disponíveis publicamente, divulgando atualizações dos dados de sequência na internet a cada 24 horas. A companhia Celera fez uso de grandes blocos de dados de sequência da IHGSC, reprocessou-os e alimentou o seu próprio banco de dados com sequências compiladas. A sequência da Celera não foi, portanto, uma sequência de genoma humano obtida independentemente. Ao contrário do IHGSC, a Celera negou acesso externo aos dados das suas sequências. Algum tempo após publicarem suas análises, a Celera requereu assinaturas a altos custos para que fosse possível visualizar os dados de sua sequência.

QUADRO 8.2 Principais marcos no mapeamento e sequenciamento do genoma humano

1956	O primeiro mapa físico do genoma humano é determinado. Utilizando microscopia óptica de tecido marcado, JH Tjio e A Levan (*Hereditas* 42, 1-6) revelaram que nossas células contêm, normalmente, 46 cromossomos e que há 24 tipos diferentes de cromossomos humanos.
1977	Fred Sanger e colaboradores publicam o método dideoxi de sequenciamento de DNA (*Proc. Natl Acad. Sci. USA* 74, 5463-5467). Com alguns refinamentos posteriores (marcação fluorescente e automação) este será o método utilizado para sequenciamento do genoma humano e muitos outros genomas.
1980	David Botstein et al. (*Am. J. Hum. Genet.* 32, 314-331) propõe que um mapa genético humano pode ser construído utilizando-se um grupo de marcadores de DNA randômicos, como polimorfismos do tamanho do fragmento de restrição (RFLPs).
1981	Sanger e colaboradores publicaram a sequência completa do DNA mitocondrial humano (Anderson S et al. *Nature* 290, 457-465).
1984	Um *workshop* é realizado em Alta, Utah, para avaliar métodos de detecção e caracterização de mutação e para projetar tecnologias futuras. Uma das conclusões principais é a necessidade de um programa de sequenciamento enorme, complexo e caro para permitir a detecção de mutações com alta eficiência.
1987	O Departamento de Energia dos Estados Unidos publica uma reportagem sobre a Iniciativa do Projeto Genoma, a qual é a primeira iniciativa desse tipo.
1987	Helen Donis-Keller e colaboradores reportam o primeiro mapa de ligação genética do genoma humano (*Cell* 51, 319-337). O mapa é baseado em RFLPs e possui uma baixa resolução.
1988	O Instituto Nacional de Saúde dos Estados Unidos (NIH) cria um departamento dedicado à Pesquisa do Genoma Humano (posteriormente renomeado para Centro Nacional para Pesquisa do Genoma Humano).
1988	A Organização do Genoma Humano (HUGO) é estabelecida para coordenar esforços internacionais, facilitando a troca de fontes de pesquisa, encorajando o debate público e alertando para as implicações da pesquisa do genoma humano.
1990	O Projeto Genoma Humano (PGH) é oficialmente lançado após a implementação de um projeto de 15 anos e de 3 bilhões de dólares nos Estados Unidos.
1992	O primeiro mapa humano abrangente de ligação genética, baseado em marcadores de microssatélites (Weissenbach J et al. *Nature* 359, 794-801).
1993	Um mapa físico de primeira geração do genoma humano é reportado, baseado em clones YAC (Cohen D et al. *Nature* 366, 698-701).
1994	Um melhorado mapa genético é publicado, baseado principalmente em marcadores de microssatélite e com um espaçamento de um centimorgan por marcador (Murray JC. et al. *Science* 265, 2049-2054).
1995	O primeiro mapa físico detalhado do genoma humano é publicado, baseado em sítios marcadores de sequências (Hudson Tj et al. *Science* 270, 1945-1954).
1996	Uma biblioteca BAC humana de alta densidade é publicada (Kim UJ et al. *Genomics* 34, 213-218).
1998	O GeneMap'98, primeiro mapa razoavelmente abrangente de marcadores baseados em genes é publicado (Deloukas P et al. *Science* 282, 744-746).
1999	A primeira sequência de DNA essencialmente completa de um cromossomo humano é reportada para o cromossomo 22 (Dunham et al. *Nature* 402, 489-495).
2001	Esboço de sequências do genoma nuclear humano, compreendendo aproximadamente 90% do componente eucromático total, é publicado pelo Consórcio Internacional de Sequenciamento do Genoma Humano (IHGSC) (*Nature* 409, 860-921) e pela Celera (*Science* 291, 1304-1351).
2001/2	Publicação de um esboço de sequências do genoma nuclear de camundongo (Waterson B et al. *Nature* 420, 520-562). Comparações entre camundongos e humanos ajudaram na identificação/caracterização dos genes humanos.
2003/4	A sequência essencialmente completa (cerca de 99%) do componente eucromático do genoma humano é reportada, compreendendo 2,85 Gb de DNA, ou aproximadamente 90% do total de 3,2 Gb (análise publicada em 2004 pelo IHGSC, *Nature* 431, 931-945).
2004	Cerca de 21 mil genes humanos são validados por clones completos de cDNA (Imanishi T et al. *PLoS Biol.* 2, e162).
2005-7	O Consórcio Internacional HapMap reporta detalhadamente mapas de polimorfismo de nucleotídeo único (SNP) para o genoma humano em 2005 (*Nature* 437, 1299-1320) e 2007 (*Nature* 449, 851-861). Este último mapa possui mais de 3,1 milhões de SNPs, aproximadamente um SNP por quilobase.
2007-8	A era do sequenciamento pessoal do genoma começa com a distribuição de sequências de genoma eucromático por James Watson e Craig Venter e o lançamento de 1.000 Projetos Genoma (http://www.1000genomes.org/page.php).

Para completar a sequência do genoma humano, o IHGSC continuou sozinho com o trabalho pesado de completar as lacunas existentes nas sequências. Em 2003, o IHGSC produziu uma sequência essencialmente completa do componente eucromático do genoma humano, a qual se tornou disponível na internet acompanhada por análises da sequência finalizada em 2004.

Os 2,85 Gb de sequências reportadas em 2003 estavam extremamente curados e arranjados em *contigs* de aproximadamente 9 Mb, em média. A sequência estava interrompida por 341 lacunas que persistiram. As lacunas são de dois tipos. Lacunas estruturais são resultantes de problemas na atribuição de mapas ambíguos, devido a sequências repetitivas (frequentemente, duplicações com sequências proximamente idênticas), mas pode ser resolvido por métodos de sequenciamento de DNA padrão (baseado em clones). Lacunas não estruturais são aquelas em que a sequência faltante não pode ser clonada em células bacterianas. Nesses casos, a sequência alterna entre nucleotídeos dos grupos pirimidina e purina, causando uma torção do DNA em uma orientação reversa, formando uma hé-

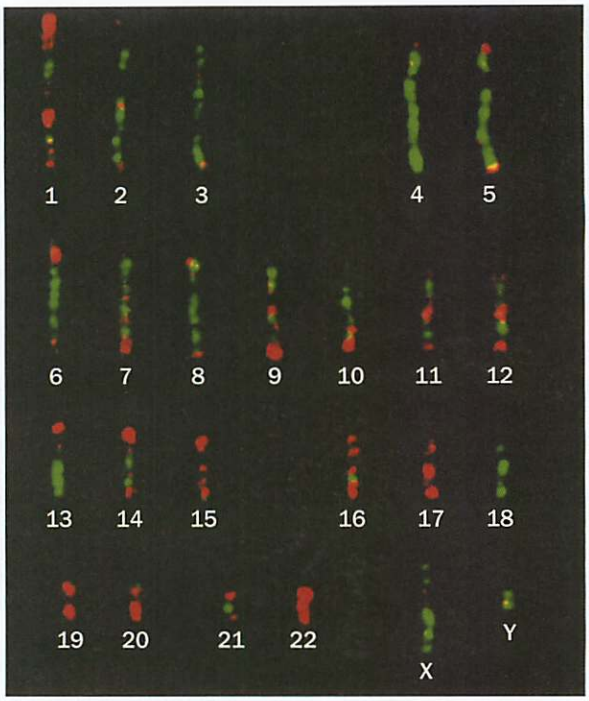

Figura 8.17 Um mapa gênico humano antigo. As extremidades 5′ da maioria dos genes humanos (e vertebrados) possuem ilhas CpG, sequências com cerca de 1 kb que diferem do volume total de DNA pela presença de dinucleotídeos CpG não metilados. A imagem mostra o resultado da hibridização de uma fração purificada de ilha CpG humana (marcada com o corante Texas Red) em cromossomos humanos em metáfase. Regiões cromossomais com replicação tardia (principalmente as transcricionalmente inativas) são distinguidas por meio da incorporação de bromodioxiuridina marcada com FITC (sinal verde). Regiões em amarelo (sobreposição dos sinais vermelho e verde) denotam regiões de replicação tardia ricas em genes (ou, estritamente, ilhas CpG). Como as ilhas CpG são marcadores gênicos, regiões cromossomais que exibem um forte sinal vermelho possuem uma alta densidade gênica (p. ex., cromossomo 22). Outros cromossomos possuem sinais em vermelho muito fracos e possuem poucos genes, como os cromossomos 4, 18, X e Y. [Adaptada de Craig JM & Bickmore WA (1994) *Nat. Genet.* 7, 376-381. Com permissão de Macmillan Publishers Ltd.]

lice com orientação para esquerda, a qual é tóxica para as bactérias. As lacunas podem ser resolvidas apenas por meio da utilização de métodos de sequenciamento por síntese. O número de lacunas no genoma eucromático tem sido reduzido progressivamente para menos de 200 na sequência mais atual (constructo 37 do NCBI; disponibilizado em 2009).

Devido à variedade de bibliotecas de DNA utilizadas para gerar a sequência, a sequência final era uma sequência composta, representando várias células doadoras com diferentes genótipos. Esta sequência agiu como uma sequência de referência que facilitou enormemente o sequenciamento subsequente dos genomas eucromáticos de indivíduos. Uma demonstração precoce foi o reporte de 2007 do sequenciamento do genoma de Jim Watson em poucos meses, utilizando pirossequenciamento massivo em paralelo. A era do sequenciamento genômico individual começou a decolar com o Projeto 1.000 Genomas, lançado no início de 2008. O objetivo é catalogar detalhadamente a variação genética humana a partir do sequenciamento dos genomas de pelo menos 1.000 indivíduos representando uma variedade de diferentes grupos étnicos (ver Quadro 8.2).

Uma vez que o esboço da sequência genômica tenha sido obtido, outro empenho internacional principal se focou em sequências funcionalmente importantes, priorizando a identificação e caracterização de todos os genes e elementos regulatórios humanos. Em 2004 foi relatado que mais de 21 mil genes humanos haviam sido validados por meio da determinação de sequências de cDNA completas. Entretanto, como descrito nas seções a seguir, ainda há incertezas consideráveis sobre exatamente quantos genes os humanos possuem.

Projetos genoma também foram conduzidos para uma variedade de organismos-modelo

No início, os objetivos do PGH incluíam o sequenciamento de cinco organismos-modelo, em parte como base para testar a tecnologia de sequenciamento. A sequência genômica de quatro dos cinco organismos priorizados pelo PGH – *E. coli, S. cerevisiae, C. elegans* e *D. melanogaster* – foi obtida durante o estágio inicial do projeto e ajudou a elaborar as estratégias de mapeamento e sequenciamento que seriam aplicadas ao genoma humano mais complexo.

Um esboço das sequências do genoma de camundongo foi obtido em 2001/2002. A companhia Celera produziu uma sequência genômica composta por múltiplas linhagens de camundongos, enquanto o Consórcio para o Sequenciamento do Genoma de Camundongo, de financiamento público, publicou a sequência genômica da linhagem de camundongo C57BL/6J, amplamente utilizada. A última atualização da sequência de C57BL/6J representa mais de 90% do total do genoma de camundongo. Como detalhado no Capítulo

10, a comparação das sequências de humanos e camundongos se provou extremamente importante na identificação de genes e na estabilização da organização éxon-íntron.

Em um estágio inicial, projetos genoma adicionais foram lançados para outros organismos, além daqueles em que o PGH havia se focado. O primeiro genoma celular foi sequenciado em 1995 (da bactéria *Haemophilus influenzae*), e, desde então, um grande número de genomas dos três reinos da vida foram sequenciados. No final de 2009, aproximadamente 1.100 sequências completas de genomas foram publicadas, e um adicional de 4.500 projetos genoma continuou em andamento. Vários genomas de arqueias e bactérias têm sido sequenciados para entender aspectos gerais e evolucionários de procariotos. A principal motivação para o sequenciamento genômico de várias bactérias tem sido o entendimento do seu envolvimento na patogênese ou em aplicações biotecnológicas. Para eucariotos, um conjunto variado de motivações estimulou o sequenciamento genômico: estes incluem modelos de pesquisa, modelos para doença e desenvolvimento, modelos para estudo genômico evolucionário e comparativo, animais de fazenda e culturas, e protozoários patogênicos e nematódeos. Dados genômicos relevantes podem ser obtidos em certas páginas da internet que fornecem compilações de bancos de dados do genoma (ver próxima seção).

A análise das sequências tem mostrado que o número de genes em um genoma é correlacionado com a complexidade de um organismo, mas a correlação é fraca. Por exemplo, foi surpreendente descobrir que *C. elegans*, um verme cilíndrico de 1 mm de diâmetro, que possui apenas 959 ou 1.031 células, dependendo do gênero, deve possuir mais de 20 mil genes codificantes, aproximadamente o mesmo número em humanos. A correlação entre conteúdo gênico e complexidade do organismo será considerada mais detalhadamente no Capítulo 10.

Bancos de dados genômicos e navegadores poderosos ajudam a estocar e analisar os dados do genoma

Logo nos estágios iniciais da análise do genoma e de genes, repositórios computacionais foram estabelecidos para estocar dados de mapeamento e dados de sequências produzidos em laboratórios por todo o mundo (**Tabela 8.3**). Após o desenvolvimento de centros de referência para sequenciamento e mapeamento genômico, um esforço em paralelo para estoque dos dados começou a partir do momento em que centros genômicos individuais desenvolveram bancos de dados próprios para estocar dados de mapeamento e sequenciamento produzidos em seus próprios laboratórios. Os dados foram disponibilizados sem custo na internet.

Conforme os dados genômicos começaram a ser produzidos em larga escala, potentes esforços foram devotados ao desenvolvimento de novos bancos de dados de genomas (**Tabela 8.4**) e ao projeto de novos *softwares* que permitissem que grandes quantidades de dados de sequência e mapeamento, assim como as informações associadas, pudessem ser procurados de maneira sistemática e acessível ao usuário. Como será visto, um foco inovador em análises *in silico* (baseado em computadores) proporcionou contribuições vitais para o entendimento da estrutura de genes e genomas.

TABELA 8.3 Principais bancos de dados eletrônicos que servem como repositórios de sequências proteicas e nucleotídicas em geral

Tipo de banco de dados	Banco de dados	Autor/hospedeiro	URL
Sequência de nucleotídeos	GenBank	Centro Nacional para Informação Biotecnológica dos Estados Unidos (NCBI)	http://www.ncbi.nlm.nih.gov
	EMBL	Instituto de Bioinformática Europeu (EBI)	http://www.ebi.ac.uk
	DDBJ	Instituto Nacional de Genética, Japão	http://www.ddbj.nig.ac.jp/
	dbEST	NCBI (divisão do GenBank que hospeda uma entrada única para cDNA/ESTs)	http://www.ncbi.nlm.nih.gov/dbEST/
Sequência de proteínas	SWISS-PROT	Instituto de Bioinformática da Suíça, Genebra	http://ca.expasy.org/sprot
		Instituto de Bioinformática Europeu	http://www.ebi.ac.uk/swissprot/
	TREMBL	EBI (tradução de sequências codificantes do banco de dados EMBL que ainda não foram depositadas no SWISS-PROT)	http://www.ebi.ac.uk/trembl/ http://ca.expasy.org/sprot
	PIR	Fundação de Pesquisa Biomédica Nacional dos Estados Unidos (NBRF)	http://pir.georgetown.edu/
		Banco de dados Internacional de Informações Proteicas do Japão (JIPID)	http://www.ddbj.nig.ac.jp/
		Centro de Informações sobre Sequências Proteicas de Munique (MIPS)	http://mips.gsf.de/

Repare que os bancos de dados possuem subdivisões que são dedicadas a classes de sequência em particular. Para sequências transcritas e genes, o banco de dados dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) e o banco de dados UniGene (http://www.ncbi.nlm.nih.gov/unigene) têm sido amplamente utilizados.

TABELA 8.4 Principais navegadores genômicos e bancos de dados de eucariotos

Fonte	Autor/hospedeiro	URL
NAVEGADORES GENÔMICOS		
Ensembl	Instituto Wellcome Trust Sanger/Instituto de Bioinformática Europeu (EBI)	http://www.ensembl.org
Visualizador de mapa NCBI	Centro Nacional para Informação Biotecnológica dos Estados Unidos (NCBI)	http://www.ncbi.nlm.nih.gov/mapview/
Navegador genômico UCSC	Universidade da Califórnia em Santa Cruz	http://genome.ucsc.edu
COMPILAÇÕES DO GENOMA		
EBI Genomas	Instituto de Bioinformática Europeu (EBI)	http://www.ebi.ac.uk/genomes
GOLD	Banco de Dados *Online* de Genomas	http://www.genomesonline.org/index.htm
BANCO DE DADOS DO GENOMA DE ORGANISMOS		
Flybase (*Drosophila*)	Consórcio *Flybase*	http://flybase.org/
MGI (*mouse genome informatics*)	Laboratório Jackson	http://www.informatics.jax.org/
NCBI *Human Genome Resources*	Centro Nacional para Informação Biotecnológica dos Estados Unidos (NCBI)	http://www.ncbi.nlm.nih.gov/projects/genome/guide/human
SGD (*Saccharomyces genome database*)	Universidade de Stanford	http://www.yeastgenome.org/
Wormbase (*C. elegans*)	Consórcio *Wormbase*	http://www.wormbase.org/
ZFIN (*Zebrafish Information Network*)	Universidade de Oregon	http://zfin.org/

Um importante avanço foi o desenvolvimento de **navegadores genômicos** com interfaces gráficas para retratar a informação do genoma para regiões individuais do cromossomo e subcromossomais. Usuários de navegadores genômicos podem navegar rapidamente pela sequência de um cromossomo humano selecionado, movendo-se de amplas escalas para escalas nucleotídicas, identificando genes e éxons associados, RNAs e proteínas (**Figura 8.18**). Conforme mais informação é obtida para genes e outras unidades funcionais, as *anotações gênicas* se tornarão mais informativas e precisas e estarão disponíveis e atualizadas com maior frequência e periodicidade nos navegadores genômicos e bancos de dados.

Diferentes programas computacionais são desenvolvidos para predizer e anotar genes dentro de sequências genômicas

Dados de sequência genômica podem ser analisados por uma série de programas computacionais, os quais procuram novos genes a partir da análise de características próprias dos genes. Por exemplo, genes são transcritos em RNA, e estes contêm sequências que mostram grande conservação evolucionária (devido à relevância de conservar importantes funções gênicas). Genes também estão associados a certos motivos e características de sequências.

Programas de predição gênica dependem fortemente da *busca por homologia* para identificar sequências evolutivamente conservadas. Uma sequência de DNA teste pode ser utilizada primeiramente na procura por sequências de RNA homólogas de qualquer organismo, utilizando programas como o BLAST (**Quadro 8.3**) para varrer bancos de dados nucleotídicos e de EST. Subsequentemente, caso se suspeite que o DNA é transcrito, ele pode ser traduzido em todas as seis fases de leitura (três para cada fita de DNA).

Os produtos da tradução predita podem, por sua vez, ser utilizados na busca por homologia com todas as sequências proteicas conhecidas (utilizando o BLASTP), assim como com produtos da tradução predita de todas as sequências nucleotídicas conhecidas (utilizando o TBLASTN). Suspeitas sequências proteicas candidatas também podem ser utilizadas como entrada na busca por homologia com bancos de dados de domínios proteicos e pequenos motivos, como detalhado no Capítulo 10.

A detecção de novos genes também pode ser auxiliada por algoritmos computacionais que reconhecem certos elementos comumente encontrados em genes. Programas de predição de éxon, como o GENSCAN, testam para sequências consenso conservadas nas junções de *splicing* (ver Figura 1.17) e atribuem alta probabilidade a um éxon predito se houver um quadro aberto de leitura que diferencia a sequência de DNA não codificante (em que, em média, um dos três códons de terminação ocorrerá a cada 60 pb).

Softwares com pacotes integrados de procura gênica têm sido desenvolvidos, os quais combinam programas desenhados para identificar éxons e motivos associados a genes

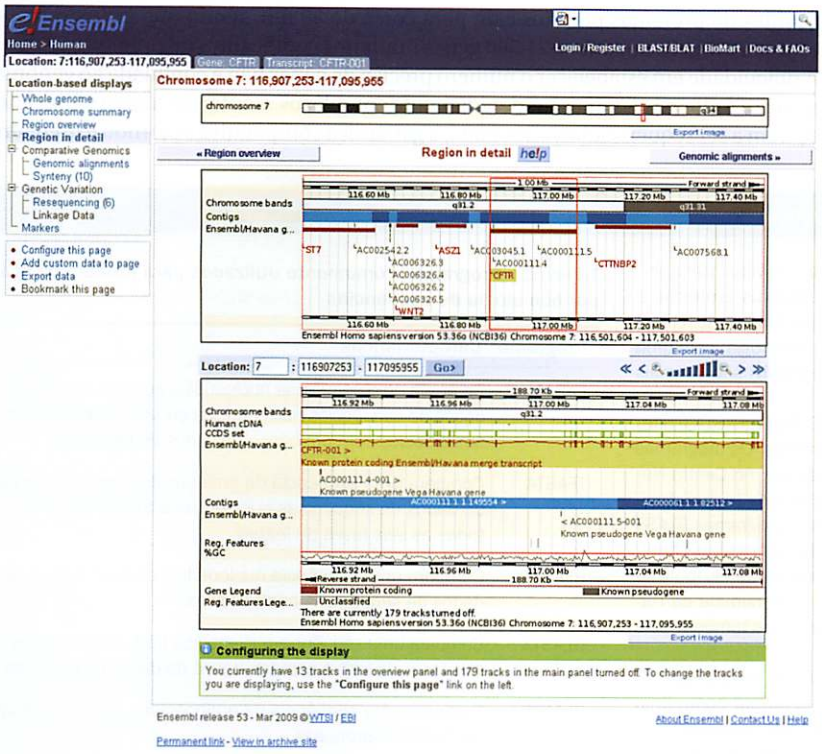

Figura 8.18 O navegador genômico Ensembl. Navegadores genômicos como o Ensembl (http://www.ensembl.org) permitem aos usuários explorar regiões subcromossomais selecionadas por meio de uma interface gráfica. Aqui, o termo de entrada foi o gene humano *CFTR* (regulador transmembrana da fibrose cística); a visão geral, na parte superior, mostra a posição do gene *CFTR* (evidenciada em verde pálido) em relação aos genes da vizinhança e marcadores de DNA no cromossomo 7q31. A visão detalhada, na parte inferior, mostra algumas das várias características que podem ser ligadas ou desligadas, de acordo com a necessidade do usuário. Na parte central, há várias ferramentas que permitem numerosas conexões à internet para outros bancos de dados e programas (inclusive conexões diretas com outras sequências montadas do genoma), permitindo aos usuários acompanhar as informações requisitadas.

com programas de busca por homologia de sequências gerais em bancos de dados. Um exemplo é o programa *Genotator* (http://www.fruitfly/org.~nomi/genotator/).

Outra característica que ajuda a identificar genes em vertebrados é a composição das bases. Genes de vertebrados estão geralmente associados com pequenas regiões, normalmente com 1 kb, que são ricas em GC e possuem uma frequência significativamente superior de dinucleotídeos CpG em relação ao volume do genoma, onde o dinucleotídeo CpG está estatisticamente sub-representado. Como essas regiões podem ser visualizadas como ilhas de frequências normal de CpG em um mar de DNA genômico, o qual, pelo contrário, é deficiente em CpG, elas são denominadas **ilhas CpG.**

Como será visto no Capítulo 9, o dinucleotídeo CpG é alvo para metilação do DNA, o qual pode causar condensação local da cromatina, inibindo a expressão gênica. As regiões à montante e outros sítios importantes que regulam a expressão gênica precisam ser capazes de adotar uma conformação mais aberta da cromatina para expressão gênica e, portanto, não são toleradas altas frequências de CpG nessa regiões. Ilhas CpG podem ser identificadas experimentalmente e a partir de programas computacionais que procuram por variação na composição das bases da sequência.

Obter estimativas acuradas para o número de genes humanos é surpreendentemente difícil

A análise da sequência do genoma humano proporcionou a identificação de milhares de genes previamente não estudados. Como será visto nos próximos capítulos, as funções dos genes preditos estão sendo estudadas intensamente na era pós sequenciamento genômico, e vocabulários sistemáticos e hierárquicos estão sendo desenvolvidos para definir as funções gênicas (**Quadro 8.4**). Surpreendentemente, entretanto, muitos anos após a sequência eucromática completa ter sido obtida, e apesar das intensas investigações, o número acurado total de genes humanos ainda não é conhecido, e levará algum tempo até que um panorama preciso seja alcançado.

Antes de o genoma humano ser sequenciado, era esperado que o número de genes humanos fosse de 70 a 100 mil, ou aproximadamente três a quatro vezes o número de genes de *C. elegans*. Em 2001, uma evidência experimental apontou cerca de 11 mil genes humanos disponíveis, mas análises do esboço da sequência genômica reportaram, naquele mesmo ano, que o número predito de genes, em sua maioria codificantes de proteínas, aumentara para apenas 30 a 40 mil. A sequência essencialmente completa do componente eucromático do genoma humano foi obtida em 2003. Após análises da nova sequência, o número de genes

codificantes de proteínas preditos caiu para cerca de 25 mil, sendo que estimativas atuais sugerem um número próximo a 21.500 genes humanos codificantes de proteínas.

A dificuldade em estabelecer o número preciso de genes não é confinada ao genoma humano, mas a todos os genomas moderadamente complexos. Conforme mais e mais genomas começaram a ser sequenciados, entretanto, a genômica comparativa se tornou extremamente

QUADRO 8.3 Busca por homologia de sequências

Programas computacionais poderosos têm sido desenvolvidos para permitir a busca rápida nos bancos de dados de sequências proteicas e nucleotídicas (e de bancos de dados de sequência genômica específica) por sequências significativamente correspondentes (homologia) com uma sequência teste. Os diferentes programas BLAST e FASTA são os mais populares (**Tabela 1**).

O desenvolvimento dos programas comparáveis BLAST e FASTA é diferente e, assim, eles podem fornecer resultados diferentes. Os programas BLAST e FASTA são amplamente disponíveis, por exemplo, no Centro Nacional para Informação Biotecnológica dos Estados Unidos (o qual também fornece tutoriais sobre como utilizar os diferentes programas BLAST (em http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html) e no Instituto de Bioinformática Europeu (http://www.ebi.ac.uk/Tools/similarityandanalysis.html); BLAT está hospedado na Universidade da Califórnia em Santa Cruz (em http://genome.ucsc.edu/), mas também está disponível em inúmeras outras localizações.

Programas como BLAST e FASTA utilizam algoritmos para identificar otimizados alinhamentos de sequência e, mostram o *output* como uma série de comparações par a par entre a sequência teste (*sequência de entrada*) e cada sequência relacionada que o programa identifica no banco de dados (*sequências-alvo*).

Diferentes abordagens podem ser empregadas para calcular o melhor alinhamento de sequências. Por exemplo, no alinhamento de sequências nucleotídicas, o algoritmo Needleman-Wunsch procura maximizar o número de nucleotídeos pareados, já o algoritmo Waterman procura minimizar o número de nucleotídeos malpareados.

Comparações par a par de alinhamentos de sequências são comparativamente simples quando as sequências teste possuem um bom pareamento e tamanhos similares, preferencialmente idênticos. Quando as duas sequências que estão sendo pareadas são significativamente diferentes uma da outra e, especialmente, quando elas possuem claras diferenças de tamanho, resultante de deleções ou inserções, é necessário um grande esforço para calcular o melhor alinhamento (**Figura 1A**).

Para sequências codificantes, o alinhamento de sequências nucleotídicas pode ser auxiliado pelo alinhamento em paralelo de sequências de aminoácidos, por meio da tradução do *frame* de leitura assumido da sequência codificante. Isto porque há 20 diferentes tipos de aminoácidos, enquanto há apenas quatro diferentes nucleotídeos. Alinhamentos par a par de sequências de aminoácidos também podem auxiliar quando se leva em consideração os grupos químicos dos aminoácidos. Substituições nucleotídicas conservativas trocam um aminoácido por outro de grupo químico relacionado, normalmente pertencendo à mesma subclasse.

Tabela 1 Programas comumente utilizados para procura básica por homologia de sequências

Programa	Características
FASTA	Compara uma sequência nucleotídica com um banco de dados de sequências nucleotídicas ou uma sequência de aminoácidos com um banco de dados de proteínas
TFASTA	Compara uma sequência de aminoácidos com um banco de dados de sequências nucleotídicas traduzidas em todas as seis fases de leitura
BLASTN	Compara uma sequência nucleotídica com um banco de dados de sequências nucleotídicas
BLASTX	Compara uma sequência nucleotídica traduzida em todas as seis fases de leitura com um banco de dados de proteínas
BLASTP	Compara uma sequência de aminoácidos com um banco de dados de proteínas
TBLASTN	Compara uma sequência de aminoácidos com um banco de dados de sequências nucleotídicas traduzidas em todas as seis fases de leitura
BLAT	Programa análogo ao BLAST que faz uma busca extremamente rápida em nível nucleotídico e proteico com um genoma sequenciado definido

Como resultado, algoritmos utilizados para comparar sequências de aminoácidos utilizam, normalmente, matrizes de escore nas quais pares de escores estão dispostos em uma matriz 20 × 20, onde altos escores são concedidos aos aminoácidos idênticos e àqueles que possuem caráter similar (p. ex., isoleucina e leucina) e baixos escores são dados aos aminoácidos que possuem caráter diferente (p. ex., isoleucina e aspartato).

Uma saída típica do programa fornece dois resultados gerais para o grau de porcentagem de relação das sequências, frequentemente denominados **porcentagem de identidade de sequência** (apenas os resíduos idênticos que parearam) e **porcentagem de similaridade de sequência** (resíduos idênticos e resíduos quimicamente relacionados que parearam; **Figura 1B**).

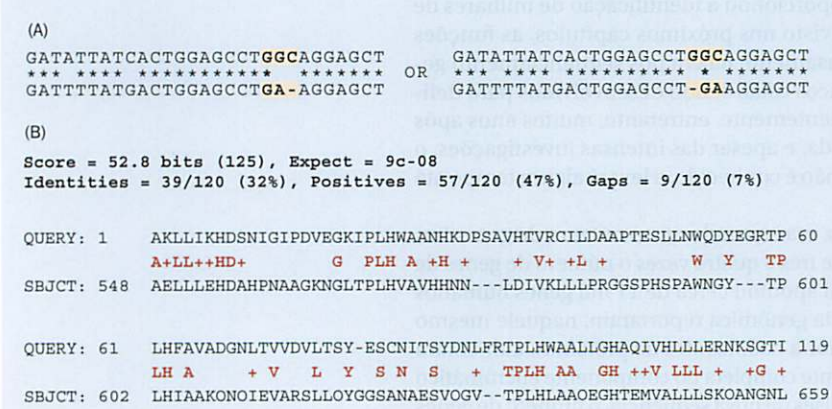

Figura 1 Alinhamento de sequências, identidade de sequências e similaridade de sequências. (A) Ambiguidade no alinhamento de sequências. As duas sequências de nucleotídeos são claramente relacionadas, mas há ambiguidade em como alinhar a sequência **GGC** com a sequência correspondente **GA**, na sequência inferior. * significa identidade; – significa ausência. (B) Identidade de sequências e similaridade de sequências. Aqui, a saída do BLASTP é o resultado da busca no banco de dados de proteínas do SWISS-PROT com a sequência de entrada de inversina. A sequência-alvo mostrada é a anquirina de eritrócitos. Aminoácidos mostrados em vermelho indicam identidade de sequência (39 de 120 posições, ou 32,5%). Há 17 posições adicionais que possuem aminoácidos quimicamente similares (mostrados como +), resultando em um total de 56 de 120 posições (46,7%) que são idênticos ou quimicamente similares.

TABELA 8.5 Dificuldades em estimar o número de genes em genomas complexos

- Predição de genes de RNA. Genes codificantes de RNAs não traduzidos podem ser difíceis de ser identificados na ausência de um quadro aberto de leitura considerável, especialmente se eles forem pequenos. Sequências de genes de RNA são, frequente e comparativamente, pouco conservadas durante a evolução. Muitas vezes é difícil distinguir entre RNAs não codificantes funcionais e sequências pseudogênicas relacionadas.
- Genes expressos a baixos níveis e/ou em localizações celulares e estágios do desenvolvimento pouco usuais podem não ser bem representados em bibliotecas de cDNA disponíveis e podem não ser evidentes se possuírem éxons pequenos.
- Genes muito grandes possuem éxons amplamente dispersos e podem ser mal interpretados como um grupo de dois ou mais genes pequenos. Este tipo de superestimativa do número de genes cresce porque muitas das bibliotecas de cDNA utilizadas para validar genes e suas organizações éxon-íntron possuem, comparativamente, pequenos insertos.
- Algumas cópias gênicas não funcionais (pseudogenes) são transcritas e podem, inicialmente, ser confundidas com genes reais.

útil na identificação de genes. Em particular, a genômica comparativa auxiliou na identificação de genes codificantes de proteínas, sendo que a estimativa recente de cerca de 21.500 genes humanos codificantes de proteínas é provavelmente bastante acurada. Genes de RNA são, entretanto, muito mais difíceis de ser identificados por duas razões principais. Primeiro, eles não possuem qualquer quadro aberto de leitura e, segundo, suas sequências tendem a ser muito menos conservadas do que as sequências codificantes de proteínas.

Durante o curso do Projeto Genoma Humano foi dada pouca atenção aos genes de RNA, sendo que várias estimativas do número de genes eram dominadas por expectativas sobre o número de genes codificantes de proteínas. Extraordinariamente, o artigo da *Science* de 2001, no qual o grupo Celera divulgou suas análises do esboço da sequência completa do genoma humano, não fazia nenhuma menção aos genes de RNA. Desde então, tem havido uma revolução no conhecimento sobre a importância dos genes de RNA. Como detalhado no Capítulo 9, aproximadamente 1.000 genes de miRNA e muitas novas classes de RNA têm sido identificadas, muitas com funções regulatórias. Pode levar algum tempo antes de se ter uma ideia clara do número de genes de RNA humanos. A **Tabela 8.5** lista algumas das dificuldades em estimar o número de genes.

8.4 ANÁLISES BÁSICAS DE EXPRESSÃO GÊNICA

Princípios para triagem da expressão

A expressão gênica pode ser monitorada tanto em nível de transcrito como de proteína com a utilização de uma variedade de tecnologias. Parâmetros importantes na expressão gênica são a fonte e a natureza do produto, da resolução, da expressão e do número de genes que são analisados por vez (**Tabela 8.6**).

Frequentemente, extratos brutos de RNA/cDNA ou proteína são utilizados como material fonte. Algumas vezes, entretanto, a expressão é amostrada em seções de tecidos ou mesmo em embriões inteiros que tenham sido fixados para preservar a morfologia original *in vivo*. A expressão também pode ser estudada em células vivas em cultura de tecidos. A expressão de genes acoplados a um grupo fluorescente também pode ser acompanhada em organismos experimentais vivos com tecidos opticamente transparentes.

A microdissecção a *laser* utiliza um *laser* para dissecar porções microscópicas de um tecido a fim de produzir populações celulares puras, a partir de fontes como biópsias teciduais e tecido marcado e até mesmo células individuais (**Figura 8.19**). Como resultado, análises de expressão gênica podem ser focadas em células únicas ou sobre populações celulares homogêneas que serão mais representativas do estado *in vivo* do que linhagens celulares.

Padrões de expressão de baixa resolução são, inicialmente, procurados em genes por meio do acompanhamento da expressão bruta de extratos de RNA e de proteína. Além de serem capazes de amostrar a expressão em diferentes tecidos, esses padrões podem fornecer informações úteis sobre o nível de expressão e sobre variantes do produto de expressão que podem diferir em tamanho (isoformas). Padrões de expressão interessantes podem ser rastreados utilizando-se métodos para acompanhar a expressão intracelular ou dentro de grupos de células e tecidos que estão organizados espacialmente de uma maneira representativa da organização *in vivo* normal.

A triagem de baixo rendimento acompanha a expressão de apenas um gene ou um pequeno número de genes por vez. Métodos de alto rendimento podem, simultanea-

QUADRO 8.4 Ontologia gênica e o Consórcio GO

Assim que o fio de água de dados de sequência genômica começou a se tornar uma inundação, biologistas começaram a lutar contra o difícil desafio de integrar dados de sequência com o vasto e rápido crescimento dos dados oriundos das análises de função gênica. A busca ao longo da ampla literatura científica e bancos de dados foi dificultada pelas grandes variações na terminologia. Era preciso um sistema padronizado de **ontologia gênica** para representar a função gênica entre genomas e espécies. O Consórcio *Gene Ontology* (GO) se formou em 1998 e começou a desenvolver um vocabulário controlado para descrever os atributos de genes e produtos gênicos em qualquer organismo, permitindo a busca por meio de múltiplos produtos gênicos e espécies por características comuns.

Para descrever os produtos gênicos, o Consórcio GO desenvolveu três ontologias em separado – processo biológico, componente celular e função molecular. Essas ontologias permitem a anotação de características moleculares ao longo de espécies e pode ser ampliada ou mais focada. Por exemplo, o processo biológico pode ser tão amplo quanto a transdução de um sinal, ou mais restrito, como no transporte de α-glicosídeo, por exemplo. Cada vocabulário é estruturado de forma que qualquer termo possa ter mais de um progenitor, bem como zero, um ou mais descendentes. O que resulta em tentativas para descrever a biologia muito mais rica do que seria possível com um gráfico hierárquico. Atualmente o vocabulário do GO consiste em mais de 17 mil termos em que todos terão, em tempo, definições estritas para seu uso. Ver http://www.geneontology.org para mais informações.

TABELA 8.6 Diferentes níveis de mapeamento da expressão

Material de estudo	Resolução	Rendimento da expressão gênica[a]	Exemplos
RNA	alta	baixa	hibridização tecidual *in situ* (Figuras 7.12 e 8.20)
			hibridização celular *in situ*
	baixa	baixa à média	hibridização por *northern blot* (Figura 7.11)
			hibridização do RNA por *dot-blot*
			ensaio de proteção à ribonuclease
			RT-PCR/qPCR
	baixa	alta	hibridização do DNA por microarranjo (Figura 8.24)
Proteína	alta	baixa	imunocitoquímica (Figura 8.22)
			microscopia por fluorescência (Figura 8.23)
	baixa	baixa	*immunoblotting* (*western blotting*) (Figura 8.21)
	baixa	alta	eletroforese em gel bidimensional (Figura 8.25)
			espectrometria de massa (Figuras 8.26 e 8.27)

[a] Número de genes ou proteínas estudado por vez.

mente, rastrear a expressão de muitos diferentes genes – geralmente centenas deles – por vez, podendo oferecer uma triagem da expressão genômica inteira.

Métodos baseados em hibridização permitem a triagem semiquantitativa e de alta resolução de transcritos de genes individuais

Uma variedade de procedimentos baseados em hibridização tem sido desenvolvida para estudar em detalhes a expressão característica de genes individuais. Estudos iniciais po-

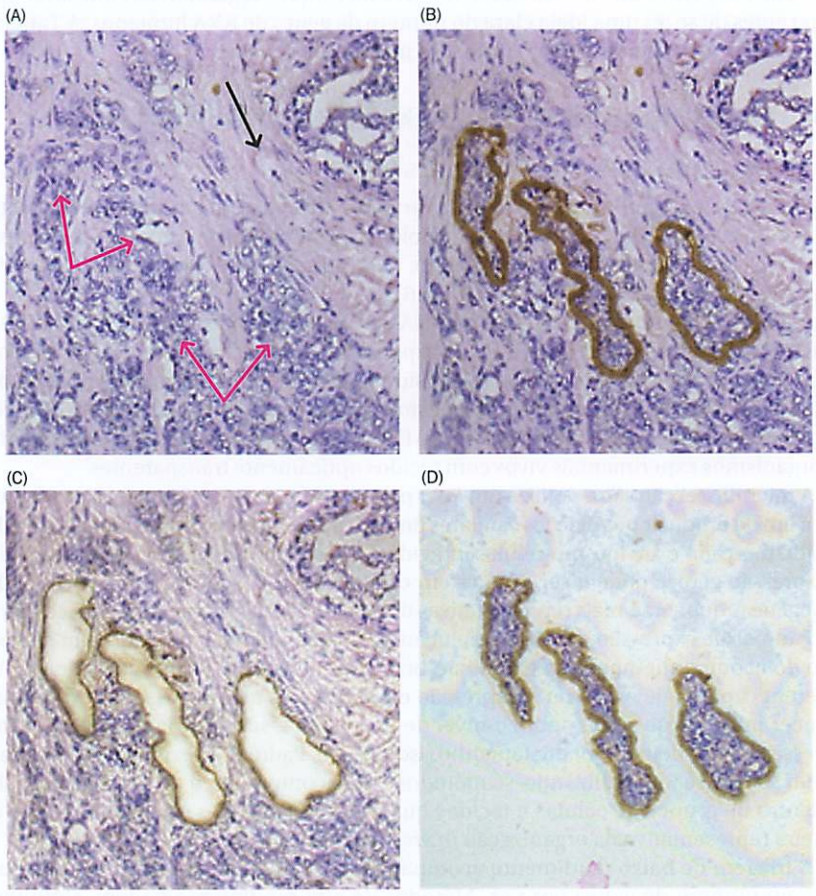

Figura 8.19 Microdissecção a *laser* de cortes teciduais. (A) Secção de próstata corada com hematoxilina e eosina. A seta em preto indica as células estromais (tecido conectivo); setas em vermelho indicam células epiteliais. (B) Contorno a *laser* das células a serem coletadas. (C) Células remanescentes após captura por *laser*. (D) Células coletadas da área contornada. [Cortesia de JR Vielkind de Garnis C, Buys TP & Lam WL (2004) *Mol. Cancer* 3, 9. Com permissão de BioMed Central Ltd.]

dem se concentrar no estabelecimento do tamanho dos transcritos de genes individuais em diferentes células e suas abundâncias aproximadas. Estudos mais avançados procurarão conduzir análises detalhadas de onde os genes são expressos em tecidos e células.

Métodos baseados em hibridização para analisar tamanho e abundância do transcrito

Na **hibridização por *northern blot***, o RNA total ou extratos de RNA poli(A)$^+$ preparados a partir de diferentes tecidos ou linhagens celulares são fracionados de acordo com o tamanho por eletroforese em gel, transferidos por *blotting* a uma membrana maleável e, então, hibridizados a uma sonda de ácido nucleico gene-específica. Esse método fornece informação sobre o tamanho e comparativa abundância de transcritos em múltiplos tecidos e linhagens celulares. Ele também pode identificar isoformas tecido-específicas que podem surgir a partir de *splicing* alternativo, poliadenilação alternativa ou uso de promotores alternativos (ver Figura 7.11).

O **ensaio de proteção à ribonuclease** procura quantificar transcritos específicos de RNA em uma complexa mistura do RNA total ou mRNA. Uma sonda de RNA fita simples antissenso marcada para o gene de interesse é incubada com uma amostra do RNA total ou mRNA para facilitar a hibridização de regiões de interesse complementares à sonda marcada. Após a hibridização, a mistura é tratada com ribonuclease (RNase), a qual digere todo o RNA fita simples, mas não digere moléculas de RNA dupla-fita. As únicas moléculas de RNA marcadas que permaneceram são aquelas que hibridizaram os transcritos de RNA específicos para formar RNA de dupla-fita. O RNA marcado é separado do RNA degradado por eletroforese em um gel de poliacrilamida desnaturante. A quantidade de detecção da marcação fornece uma medida da quantidade de transcrito específico na amostra.

Hibridização tecidual *in situ*

Padrões de expressão espacial de RNA em alta resolução em tecidos e grupos de células são normalmente obtidos por **hibridização tecidual *in situ***. Normalmente, tecidos são congelados ou embebidos em cera, e então cortados com um micrótomo para fornecer finas fatias (p. ex., 5 μm ou menos), as quais são montadas em uma lâmina de microscópio. A hibridização de uma sonda gene-específica adequada ao tecido na lâmina pode fornecer imagens da expressão detalhadas, representativas da distribuição do RNA no tecido de origem. Tecidos embrionários são muito utilizados: seu pequeno tamanho permite a triagem de vários tecidos em uma única seção (ver Figura 7.12).

Uma extensão da hibridização tecidual *in situ* é o estudo da expressão em um embrião intacto. A hibridização *in situ whole mount* é um método popular para acompanhar a expressão durante o desenvolvimento de embriões inteiros a partir de organismos vertebrados modelos. Por causa das dificuldades éticas e práticas em conduzir análises gênicas equivalentes em humanos, há considerável confiança na extrapolação de análises conduzidas sobre embriões de camundongos ou aquelas sobre outros modelos de vertebrados (**Figura 8.20**). A quantidade relativamente grande de tecido disponível mostra que o método é relativamente sensível e, além disso, a automatização da técnica tem aumentado a sua popularidade.

Com a utilização de sondas adequadas, sequências de RNA específicas também podem ser rastreadas dentro de *células individuais* para identificar sítios de processamento do RNA, transporte e localização citoplasmática. Com o uso de hibridização *in situ* fluorescente quantitativa (FISH) e microscopia de imagem digital, tem sido possível detectar até mesmo transcritos de RNA *in situ*. Um refinamento posterior utiliza combinações de diferentes tipos de sondas oligonucleotídicas marcadas com fluoróforos de diferentes espectros. Isso tem permitido que transcritos de múltiplos genes sejam acompanhados simultaneamente.

Figura 8.20 Hibridização *in situ whole mount*. O exemplo mostra a expressão do gene do fator de crescimento de fibroblasto, *Fgf8*, no embrião de galinha no estágio 20 de Hamburger Hamilton (cerca de 3 dias do desenvolvimento após a postura do ovo). Transcritos foram marcados com uma sonda de *Fgf8* antissenso marcada com digoxigenina e foram detectados com anticorpos antidigoxigenina acoplados a fosfatase alcalina. O ensaio de fosfatase alcalina utilizou uma combinação de BCIP (5-bromo-4-cloro-3-indolil-fosfato) e NBT (Nitro Blue Tetrazolium), resultando em fortes sinais de expressão azuis. A expressão foi evidente no desenvolvimento do olho, istmo, arcos branquiais, somitos, brotamento dos membros superiores e da cauda. (Cortesia de Terence Gordon Smith, Newcastle University.)

Métodos por PCR quantitativo são amplamente utilizados na triagem da expressão

Os métodos descritos para triagem da expressão baseados em hibridização são desenvolvidos para detectar transcritos que não foram amplificados. Sendo assim, não são adequados para a detecção de transcritos com baixo número de cópias. Métodos baseados em PCR são muito mais sensíveis e podem rastrear a expressão gênica em tipos celulares ou tecidos que não são fáceis de ser acessados em grande quantidade, ou até mesmo em células individuais. Há uma variedade de tais métodos, mas todos requerem que o RNA de início seja primeiramente transformado em uma sequência de cDNA pela utilização da transcriptase reversa.

No método básico de **PCR com transcriptase reversa** (**RT-PCR**), uma cópia de cDNA é primeiramente produzida a partir de um RNA utilizando-se um iniciador (*primer*) oligo(dT) ou iniciadores oligonucleotídicos de sequência randômica, e estes são utilizados para iniciar a reação de PCR. Depois de a reação de PCR estar completa, uma alíquota é obtida para análise em eletroforese padrão em gel para fracionar o cDNA de acordo com o tamanho, onde então os produtos separados se ligam a brometo de etídio e são revelados sob radiação ultravioleta. O método básico de RT-PCR tem sido útil para a identificação e o estudo de diferentes isoformas de um transcrito de RNA. Entretanto, a detecção do DNA com fluorescência pelo brometo de etídio não é muito sensível. Ele pode detectar produtos que tenham passado por 30-40 ciclos completos de PCR, mas nesse momento o estágio exponencial de amplificação (a duplicação regular do produto por ciclo de PCR) já passou e não é mais possível obter uma quantificação acurada de RNA.

Para obter uma quantificação acurada de transcritos de RNA, um método variante é utilizado, o qual foi inicialmente chamado de **PCR em tempo real**, mas que agora é simplesmente chamado **qPCR** (PCR quantitativo). Ao contrário do método básico de RT-PCR, o qPCR não envolve o fracionamento por eletroforese em gel dos produtos amplificados finais. Porém, o qPCR é uma reação *cinética* que requer equipamento de PCR especializado. No qPCR os produtos da amplificação são quantificados simultaneamente durante a reação de PCR. Para obter dados mais acurados, as medidas de quantificação são feitas somente durante os estágios iniciais da reação de PCR, quando a amplificação ainda é exponencial.

Para detectar produtos de amplificação durante uma reação de qPCR, alguns sinais mensuráveis devem ser gerados, os quais são proporcionais à quantidade de produto amplificado. Todos os métodos de detecção atuais utilizam tecnologias de fluorescência, sendo que o método de detecção pode ser específico ou não específico (**Quadro 8.5**).

Anticorpos específicos podem ser utilizados para rastrear proteínas expressas por genes individuais

Devido a sua extraordinária diversidade, seletividade e sensibilidade na detecção de proteínas, anticorpos são ideais para rastrear a expressão gênica a nível proteico. A maneira tradicional para obter um anticorpo é injetar repetidamente em um animal adequado um **imunógeno** específico, uma molécula que é detectada pelo sistema imune do hospedeiro como sendo externa. Para proliferar um anticorpo contra uma proteína específica, o imunógeno utilizado é um peptídeo sintético ou uma proteína de fusão que contenha parte da sequência proteica que se quer rastrear (**Quadro 8.6**).

Anticorpos podem ser marcados de diferentes maneiras. Nos métodos de detecção direta, o anticorpo purificado é marcado a partir do acoplamento de uma molécula repórter, como um fluoróforo ou biotina, permitindo que o anticorpo marcado se ligue diretamente à proteína-alvo. Alternativamente, a proteína-alvo é ligada primeiramente por um *anticorpo primário* não marcado que se liga à proteína-alvo, o qual é então ligado especificamente por uma molécula secundária marcada que pode ser um *anticorpo secundário*, ou seja, um anticorpo específico que reconhece o anticorpo primário. Algumas vezes, uma molécula geral é utilizada, como a proteína A, encontrada na parece celular de *Staphylococcus aureus*. Por razões desconhecidas, a proteína A se liga fortemente a uma região central comum em moléculas de anticorpo (na segunda e terceira regiões constantes da porção Fc das cadeias pesadas de imunoglobulina).

Anticorpos marcados podem ser utilizados por meio de vários métodos para acompanhar proteínas específicas em extratos livres de células, tecidos fixados e células. Para extratos livres de células, uma aplicação comum é o ***immunoblotting*** (***western blotting***). Neste método, proteínas de extratos livres de células são primeiramente dissolvidas em uma solução com dodecil sulfato de sódio (SDS), um detergente aniônico que quebra aproximadamente todas as interações não covalentes em proteínas nativas e então separadas de acordo com o tamanho por eletroforese em gel de poliacrilamida-SDS (SDS-PAGE). As proteínas fracionadas são transferidas (*blotted*) para uma membrana de nitrocelulose e então expostas a um anticorpo específico (**Figura 8.21**). Como descrito na Seção 8.5, a eletroforese bidimensional em gel de poliacrilamida também pode ser utilizada.

Na imunocitoquímica (= imunohistoquímica), um anticorpo é utilizado para obter o padrão de expressão geral para uma proteína dentro de um tecido ou de outra estrutura multicelular. Assim como na hibridização *in situ*, os tecidos são, geralmente, ou congelados ou embebidos em cera e, então, cortados em secções finas com um micrótomo antes de serem montados em uma lâmina. Um anticorpo específico adequado se liga à proteína

QUADRO 8.5 Métodos de detecção utilizados no PCR quantitativo (em tempo real)

A quantificação no qPCR depende da detecção do sinal de fluorescência que é gerado após a ligação de algum reagente ao produto amplificado nos estágios iniciais da reação de PCR. Métodos de detecção não específica são limitados à detecção de apenas um tipo de amplicon-alvo por vez, enquanto métodos de detecção específica podem distinguir entre diferentes alvos e permitem ensaios múltiplos.

A detecção específica é possibilitada pelo uso de sondas de hibridização de fita simples que são desenhadas para se ligar a um tipo específico de amplicon durante o estágio de anelamento da reação de qPCR. Como detalhado na tabela e figuras a seguir, métodos populares de detecção específica utilizam, frequentemente, sondas de hibridização que carregam um fluoróforo na extremidade 5′ e um grupo *quencher* na extremidade 3′ que absorve fótons emitidos pelo fluoróforo e, então, dissipa a energia absorvida ou na forma de calor ou na forma de luz com diferentes comprimentos de onda.

Método	Base do método
Detecção não específica	
Corante SYBR Verde I	Quando livre em solução, este corante mostra relativamente pouca fluorescência, mas quando se liga à dupla-fita do DNA, sua fluorescência aumenta mais de 1.000 vezes. No entanto, a sua ligação ao DNA dupla-fita não é específica, podendo ocorrer a sua ligação a iniciadores (*primers*) diméricos ou produtos não específicos da amplificação (por isso não sendo utilizado na PCR padrão). Para controlar isso, uma *curva de dissociação* é conduzida no fim do processo, aumentando gradualmente a temperatura de 60°C para 95°C, enquanto a fluorescência é monitorada continuamente. A uma certa temperatura, todo o produto amplificado irá se dissociar completamente, resultando em um decréscimo da fluorescência conforme o corante se dissocia do DNA. A temperatura de dissociação é dependente do tamanho e da composição do amplicon, permitindo que diferentes fragmentos de DNA sejam distinguidos (deve haver amplificação não específica ou iniciadores diméricos significativos).
Detecção específica a partir de sondas de hibridização	
Sondas Molecular Beacon	Sondas Molecular Beacon são desenhadas para ter uma estrutura na forma haste-alça que carrega um fluoróforo na extremidade 5′ muito próximo a um *quencher* na extremidade 3′, limitando de modo severo a fluorescência emitida pelo fluoróforo. Entretanto, na presença de uma sequência-alvo complementar, a sonda se desdobra e hibridiza com o alvo, deslocando o fluoróforo para longe do *quencher* (**Figura 1**). Como resultado, o *quencher* não absorve mais os fótons emitidos pelo fluoróforo e a sonda começa a emitir fluorescência.
Sondas TaqMan® com dupla marcação	Aqui a sonda é um oligonucleotídeo que possui um fluoróforo, como o FAM, na extremidade 5′, e um grupo *quencher*, como o TAMRA ou o *Black Hole Quencher*, na extremidade 3′. Nesse caso, o mecanismo de absorção é baseado na transferência de energia de ressonância por fluorescência (FRET, do inglês *Fluorescence Resonance Energy Transfer*) e pode ocorrer sob uma distância relativamente grande – há 10 nm ou mais. Como mostrado anteriormente, a sonda oligonucleotídica pode se ligar à sequência complementar do amplicon do qPCR durante o passo de anelamento da reação de PCR, mas nesse caso o fluoróforo da sonda hibridizada intacta não emite fluorescência porque continua transmitindo sua energia para o *quencher* (**Figura 2**). Após a ligação da sonda ao amplicon alvo, entretanto, o avanço da Taq-polimerase desloca a extremidade 5′ da sonda, a qual é então degradada pela atividade de exonuclease 5′→3′ da Taq-polimerase. Enquanto a polimerase desloca o resto da sonda, a clivagem continua, resultando na liberação do fluoróforo e do *quencher* em solução (**Figura 3**). Agora que eles estão fisicamente separados um do outro, o fluoróforo FAM exibe forte fluorescência, mas o *quencher* emite muito menos energia (em um comprimento de onda diferente daquele do fluoróforo).

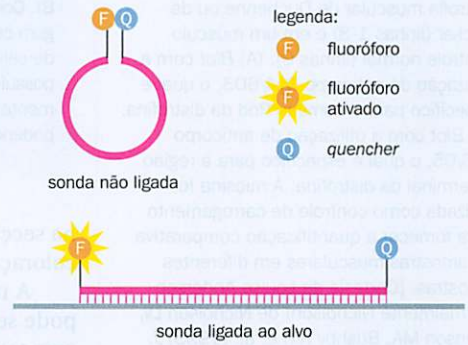

Figura 1 Sondas Molecular Beacon fluorescem apenas após terem sido hibridizadas ao DNA-alvo.

Figura 2 Hibridização da sonda de oligonucleotídeo TaqMan® a um amplicon de PCR. F, fluoróforo; Q, *quencher*.

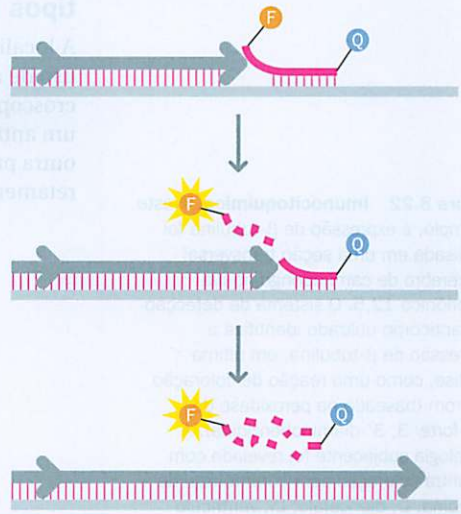

Figura 3 A sonda TaqMan® hibridizada é deslocada pela Taq-polimerase e degradada pela exonuclease associada, ativando assim o fluoróforo da sonda.

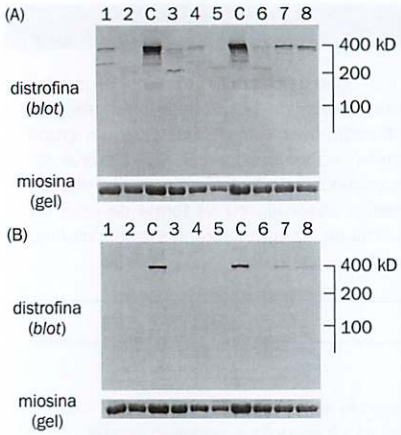

Figura 8.21 ***Imunoblotting* (*western blotting*).** O *imunoblotting* envolve a detecção de polipetídeos após o fracionamento por tamanho em gel de poliacrilamida e transferência (*blotting*) para uma membrana. Aqui, os *blots* mostram a detecção de distrofina em amostras musculares de pacientes individuais com distrofia muscular de Duchenne ou de Becker (linhas 1-8) e em um músculo controle normal (linhas C). (A) *Blot* com a utilização de anticorpo Dy4/6D3, o qual é específico para o domínio Rod da distrofina. (B) *Blot* com a utilização de anticorpo Dy6/C5, o qual é específico para a região C-terminal da distrofina. A miosina foi utilizada como controle de carregamento para fornecer a quantificação comparativa de amostras musculares em diferentes amostras. [Cortesia de Louise Anderson (formalmente Nicholson) de Nicholson LV, Johnson MA, Bushby KM et al. (1993) *J. Med. Genet.* 30, 737-744. Com permissão de BMU Publishing Group.]

QUADRO 8.6 Obtendo anticorpos

Anticorpos que podem detectar especificamente uma proteína de interesse podem ser proliferados de diferentes maneiras. Uma abordagem comparativa recente tem sido a utilização da técnica de *exposição em fagos* (*phage display*), um sistema de clonagem de expressão no qual fagos recombinantes são utilizados para expor proteínas heterólogas na superfície de células bacterianas (ver Seção 6.3). Entretanto, o principal método para obtenção de anticorpos tem sido a rota tradicional de administração repetida em animais (como roedores, coelhos e cabras) com um imunógeno que representa a proteína sob investigação. Uma abordagem é desenhar um peptídeo sintético, frequentemente em torno de 20 a 50 aminoácidos, o qual é então conjugado com uma molécula adequada (como a hemocianina de *Megathura crenulata*) que ajuda a maximizar a imunogenicidade. A esperança é que o peptídeo adotará uma conformação similar à sequência polipeptídica nativa, mas o sucesso não é garantido, e pode ser necessário que vários peptídeos diferentes sejam desenhados.

Um tipo alternativo de imunógeno é uma **proteína de fusão** que contém a maioria, ou uma grande porção, da sequência proteica alvo fusionada a outra proteína que confere algumas vantagens, principalmente auxiliando na purificação proteica. A proteína de fusão é sintetizada por meio da clonagem de um cDNA com uma proteína-alvo desejada em um vetor plasmidial de expressão que também contenha um cDNA para a proteína associada e para a seleção de recombinantes onde os dois cDNAs estejam na mesma fase de leitura (ver Figura 6.12).

Se o sistema imune do animal reagiu, anticorpos específicos devem ser secretados no soro. O soro rico em anticorpos (antissoro) contém uma mistura heterogênea de anticorpos, cada um produzido a partir de um linfócito B diferente. Os diferentes anticorpos reconhecem diferentes partes (*epitopos*) do imunógeno e são conhecidos como **anticorpos policlonais**.

Uma preparação homogênea de anticorpos com uma especificidade definida pode ser preparada, entretanto, por meio da propagação de um clone de células (originalmente oriunda de um único linfócito B). Como as células B possuem um tempo de vida em cultura limitado, é preferível estabelecer uma linhagem celular imortal: células produtoras de anticorpos são fusionadas com células derivadas de um tumor de célula B imortal. A partir da mistura heterogênea resultante de células híbridas, aqueles híbridos que possuírem a habilidade de produzir um anticorpo em particular e a habilidade de se multiplicar indefinidamente em cultura são selecionados. Tais **hibridomas** são propagados como clones individuais, cada qual podendo fornecer uma fonte permanente e estável de um único tipo de **anticorpo monoclonal** (**mAb**).

na secção do tecido e pode produzir dados de expressão que podem estar relacionados à coloração histológica das secções teciduais vizinhas (**Figura 8.22**).

A microscopia eletrônica fornece resoluções maiores do que a imunocitoquímica e pode ser utilizada para investigar a localização intracelular de uma proteína em células com tecido fixado que tenham sido cortadas em secções ultrafinas. O anticorpo é normalmente marcado com uma partícula de densidade eletrônica, como esferas coloidais de ouro. Como descrito na próxima seção, diferentes métodos de microscopia por fluorescência são utilizados para rastrear a expressão subcelular de proteínas de células em cultura.

Expressão proteica em cultura de células é frequentemente analisada pelo uso de diferentes tipos de microscopia de fluorescência

A localização subcelular de uma proteína de interesse também pode ser rastreada com o uso de anticorpos em cultura de células, por meio da utilização de diferentes tipos de microscopia fluorescente. A proteína pode ser acompanhada diretamente com a utilização de um anticorpo específico proliferado contra ela. Alternativamente, uma cauda peptídica ou outra proteína é acoplada à proteína de interesse, e a expressão da proteína é seguida indiretamente. Na microscopia por imunofluorescência, anticorpos rastreiam a proteína direta-

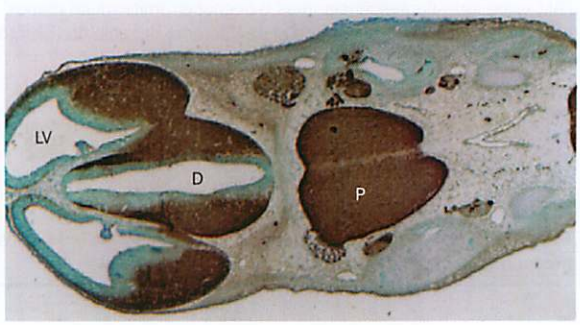

Figura 8.22 Imunocitoquímica. Neste exemplo, a expressão de β-tubulina foi analisada em uma seção transversal do cérebro de camundongo no dia embriônico 12,5. O sistema de detecção por anticorpo utilizado identifica a expressão de β-tubulina, em última análise, como uma reação de coloração marrom (baseada na peroxidase de raiz-forte/ 3, 3′-diaminobenzidina). A histologia subjacente foi revelada com a contra coloração com corante azul de toluidina. D, diencéfalo; LV, ventrículo lateral; P, bulbo. (Cortesia de Steve Lisgo, Newcastle University.)

TABELA 8.7 Epitopos marcadores comumente utilizados para monitorar localização proteica

Sequência marcadora	Peptídeo de origem	Ligado à proteína em	Anticorpo monoclonal
DYKDDDDK	sintético	N- ou C- terminal	anti-Flag M1
EQKLISEEDL	proteína humana c-Myc	N- ou C- terminal	9E10
MASMTGGQQMG	proteína do fago T7 gene 10	N-terminal	anticorpo T7.Tag
QPELAPEDPED	proteína D de vírus herpes *simplex*	C-terminal	anticorpo HSV.Tag
RPKPQQFFGLM	substância P	C-terminal	NC1/34
YPYDVPDYA	proteína de *influenza* HA1	N- ou C- terminal	12CA5

mente. Os anticorpos são marcados com um marcador fluorescente adequado, como fluoresceína ou rodamina, e após a ligação do anticorpo à proteína de interesse, a expressão da proteína é monitorada diretamente dentro das células com o uso de microscopia por fluorescência (ver Figura 2 do Quadro 7.3 para o princípio da microscopia por fluorescência).

Na **marcação de epitopos (*epitope tagging*)**, a proteína relevante é marcada por meio do ancoramento de uma sequência peptídica imunogênica marcadora à extremidade de N- ou à C-terminal, para a qual um anticorpo específico já exista. Para isso, um cDNA para a proteína de interesse é clonado em um vetor de expressão em um sítio adjacente a, e na mesma fase de leitura com, uma sequência codificante para o peptídeo de relevância. Plasmídeos recombinantes são transfectados nas células apropriadas e a proteína expressa é monitorada com um anticorpo marcado com fluorescência específico para o epitopo marcador. Epitopos marcadores comumente utilizados são mostrados na **Tabela 8.7**.

Mais recentemente, a microscopia por fluorescência tem sido utilizada para monitorar a expressão de proteínas em células sem a utilização de anticorpos. Uma proteína naturalmente fluorescente é acoplada à proteína de interesse para agir como marcador do seu padrão de expressão. Uma dessas proteínas é a proteína verde fluorescente (GFP), uma proteína de 238 aminoácidos originalmente identificada na água-viva *Aequoria victoria*. Proteínas similares são expressas em várias águas-vivas e parecem ser responsáveis pela luz verde que elas emitem, sendo estimuladas pela energia obtida após a oxidação da luciferina ou outra fotoproteína.

Quando o gene codificante da GFP foi clonado e transfectado para células-alvo em cultura, a expressão de GFP em células heterólogas também foi marcada pela emissão de luz fluorescente verde. Isso significa que a GFP é uma proteína *autofluorescente*. Devido à sua propriedade natural de fluoróforo funcional, ela pode servir diretamente como um repórter que pode ser seguido prontamente por microscopias de fluorescência convencional e confocal, fazendo destas ferramentas populares para monitoramento da expressão gênica em linhagens celulares e, até mesmo, em alguns organismos inteiros.

Vários mutantes de GFP geneticamente modificados melhoraram a sua eficiência de diferentes maneiras, e vários mutantes coloridos têm sido gerados, especialmente proteína azul fluorescente, proteína ciano fluorescente e proteína amarelo fluorescente. Outra proteína autofluorescente, a proteína vermelha fluorescente, tem sido obtida a partir do coral *Discosoma*.

Para monitorar a expressão de uma proteína dentro de células vivas, a sequência de DNA codificante de GFP tem sido inserida no início ou no fim do cDNA para a proteína de interesse dentro de um vetor de expressão adequado. A transfecção e a expressão da sequência de cDNA híbrida resultante produz uma proteína de fusão com GFP ligado à extremidade N- ou C-terminal da proteína de interesse. Em muitos casos, a proteína de fusão é expressa da mesma maneira que a proteína original, ajudando a revelar a sua localização (**Figura 8.23**).

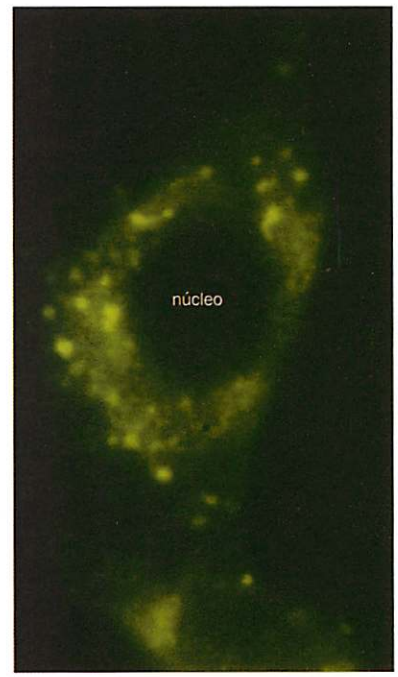

Figura 8.23 Acompanhamento da expressão proteica utilizando-se cauda de proteína verde fluorescente. Este exemplo mostra uma célula viva transiente de HeLa transfectada expressando a proteína da doença de Batten marcada com GFP. Uma sequência cDNA do gene *CLN3* da doença de Batten foi clonada no vetor de expressão de GFP pEGFP-N1 para expressar a proteína CLN3 com a sequência de GFP acoplada à região C-terminal (uma *cauda GFP*). Essa célula é um exemplo de uma pequena proporção de células HeLa expressando CLN3/GFP em um padrão vesicular de pontos distribuídos pelo citoplasma. Estas e outras análises indicaram que a proteína da doença de Batten é uma proteína integral da membrana de Golgi. [De Kremmidiotis G, Lensink IL, Bilton RL et al. (1999) *Hum. Mol. Genet.* 8, 523-531. Com permissão de Oxford University Press.]

8.5 ANÁLISES AVANÇADAS EM PARALELO DA EXPRESSÃO GÊNICA

Análises avançadas em paralelo da expressão gênica permitem a triagem simultânea da expressão de centenas ou milhares de genes, até mesmo o grupo total de diferentes transcritos (transcriptoma) ou das diferentes proteínas (**proteoma**) expressos pela célula. Para o perfil transcricional em larga escala, a hibridização por microarranjo tem sido o método mais utilizado; entretanto, métodos alternativos baseados em sequenciamento vêm sendo desenvolvidos, sendo que o sequenciamento massivo em paralelo está começando

a revolucionar as análises do transcriptoma. O perfil proteico global tem sido alcançado popularmente com eletroforese em gel bidimensional e espectrometria de massa.

Microarranjos de oligonucleotídeos e DNA permitem a rápida obtenção do perfil global de transcritos

Os alvos para traçar o perfil de transcritos são complexas populações de RNA de fontes celulares de interesse, geralmente cultura de células, tecidos ou tumores excisados cirurgicamente ou porções isoladas destes. Microarranjos típicos utilizam centenas ou milhares de sondas gene-específicas. Sondas de cDNA têm sido utilizadas, mas sondas oligonucleotídicas estão sendo cada vez mais empregadas. Como detalhado na Seção 7.4, dois sistemas populares são os microarranjos Affymetrix GeneChip, no qual oligonucleotídeos com cerca de 25 nucleotídeos são sintetizados *in situ* na própria lâmina, e os microarranjos Illumina, em que oligonucleotídeos pré-sintetizados de cerca de 50 nucleotídeos com um componente gene-específico são acoplados a esferas. Mais recentemente, arranjos de alta densidade com oligonucleotídeos maiores com até 60 nucleotídeos também estão se tornando disponíveis.

Análises de expressão baseadas em microarranjo são geralmente organizadas visando a comparação de duas ou mais fontes celulares ou teciduais relacionadas que diferem de uma maneira informativa. Por exemplo, a expressão gênica pode ser acompanhada em estruturas embrionárias específicas em uma série de etapas do desenvolvimento, ou a expressão pode ser monitorada em culturas celulares em tempos diferentes, após serem expostas a uma série de drogas e a outras substâncias químicas. O mesmo tipo de tecido tumoral pode ser caracterizado em diferentes estágios de malignidade. Perfis de expressão para fenótipos doentes e normais também podem ser comparados, sendo que em modelos de camundongos as duas fontes celulares sob comparação podem ser genotipicamente idênticas, exceto pela presença de uma mutação patogênica definida.

Para realizar a caracterização de transcrito, a amostra de RNA celular é, normalmente, transcrita reversamente em massa para formar uma população de cDNAs complexos representativos. O cDNA pode ser marcado à medida que é sintetizado, por meio da inclusão de um nucleotídeo conjugado a um fluoróforo na mistura da reação. Alternativamente, um procedimento de dois passos é utilizado – primeiro, o cDNA não marcado é construído, para então ser convertido em um RNA complementar (cRNA) marcado, por meio da incorporação de biotina (a qual será detectada posteriormente com estreptavidina conjugada a um fluoróforo).

O cDNA ou cRNA alvo marcado é então aplicado sobre a lâmina para hibridizar. Cada característica individual, ou ponto, sobre a lâmina contém de milhões a bilhões de cópias da mesma sequência de DNA, sendo, portanto, improvável ocorrer saturação completa na reação de hibridização. Sob essas condições, a intensidade do sinal de hibridização sobre cada característica sobre a lâmina é proporcional à relativa abundância daquele cDNA ou cRNA em particular da população alvo, refletindo, consequentemente, a abundância do mRNA correspondente à população fonte original. A relativa abundância de milhares de diferentes transcritos pode, portanto, ser monitorada em um experimento. Múltiplos oligonucleotídeos também são utilizados para ajudar a distinguir entre transcritos fortemente relacionados a genes individuais. É possível monitorar variantes de *splicing*, por exemplo, e desenvolver oligonucleotídeos específicos para cada éxon conhecido.

A grande quantidade de dados de expressão gerados pelos dados de microarranjo requer uma análise estatística cuidadosa (**Quadro 8.7**). Controles estringentes também são requeridos para normalizar os dados de expressão para variação entre experimentos. Uma maneira de evitar tais problemas é hibridizar populações de cDNA marcadas com diferentes fluoróforos na mesma lâmina, simultaneamente. Sob condições de não saturação, o sinal de cada característica representará a relativa abundância de cada transcrito na amostra. Se duas amostras são utilizadas, a taxa de sinais de cada fluoróforo fornece uma comparação direta dos níveis de expressão entre amostras totalmente normalizadas para variações na taxa do sinal pelo ruído, mesmo quando dentro do arranjo. A lâmina é lida em dois comprimentos de onda e um computador é utilizado para combinar as imagens e revelá-las em falsa cor. Normalmente, um fluoróforo é representado como verde e outro como vermelho. Características que representam genes expressos diferencialmente são mostradas ou em verde ou em vermelho, enquanto aquelas que representam genes equivalentemente expressos são mostradas em amarelo – ver exemplo do uso de pontos de cDNA em lâminas na **Figura 8.24A**.

As análises da expressão com microarranjos que utilizam sondas oligonucleotídicas pequenas é similar, a princípio, àquelas nos quais sondas de cDNA ou oligonucleotídeos maiores (mais de 50 nucleotídeos) são utilizados. Entretanto, ao utilizar oligonucleotídeos que possuem apenas 25 nucleotídeos, a especificidade da reação não é tão grande.

QUADRO 8.7 Analisando dados de expressão do microarranjo

Dados de expressão do microarranjo estão revolucionando as pesquisas biomédica e biológica, mas a interpretação dos dados ainda é um desafio. Os dados de expressão brutos oriundos de experimentos de microarranjo são sinais de diferentes intensidades que devem ser corrigidos de acordo com os efeitos de fundo e variação entre experimentos (*normalização*) e conferidos para erros causados por contaminantes e valores extremos afastados.

Os dados são resumidos como uma tabela de sinais normalizados com diferentes intensidades: as linhas da tabela representam genes individuais, e as colunas, diferentes condições sob as quais a expressão gênica foi mensurada. Nos casos mais simples, a tabela possui duas colunas (p. ex., amostras controle e doente) e estas podem representar a intensidade de sinal de duas amostras hibridizadas simultaneamente sobre a lâmina. Entretanto, teoricamente não há limite para o número de condições que podem ser utilizados.

Após, genes com perfis de expressão similar são agrupados. Geralmente, quanto mais as condições sob as quais a expressão gênica é testada mais rigorosa é a análise. Dois tipos de algoritmos são utilizados para minerar os dados de expressão gênica: um em que os dados similares são agrupados hierarquicamente e outro em que os grupos são definidos de forma não hierárquica.

Agrupamento hierárquico

A abordagem geral no agrupamento hierárquico é estabelecer uma *matriz de distância* que lista os diferentes níveis de expressão entre cada par de características sobre a lâmina. Aqueles que exibem as menores diferenças, expressos como a função de distância *d*, são então agrupados de uma maneira progressiva. Métodos de *agrupamento aglomerativo* começam com a classificação de cada gene representado sobre a lâmina como um grupo singular (um grupo contendo um gene). A matriz de distância é analisada e os dois genes com os níveis de expressão mais similares (a menor matriz de distância) são definidos como vizinhos; estes são então unidos em um único grupo. O processo é repetido até que sobre apenas um grupo. Com a proposta de futuras comparações, há variações no modo como o valor de expressão do grupo unido é calculado.

No método dos *vizinhos mais próximos* (*ligação única*) a distância é minimizada. Ou seja, onde dois genes *i* e *j* são unidos em um único grupo *ij*, a distância entre *ij* e o próximo gene vizinho *k* é definida como o menor dos dois valores *d*(*i*,*k*) e d(*j*,*k*). No *método da ligação única média*, a média entre d(*i*,*k*) e d(*j*,*k*) é utilizada. No método dos *vizinhos mais afastados* (*ligação completa*), a distância é maximizada. Estes métodos geram dendogramas com diferentes estruturas (**Figura 1**). Com menor frequência, um algoritmo de divisão de grupos pode ser utilizado, em que um único grupo representando todos os genes sobre a lâmina é progressivamente separado em grupos distintos.

O agrupamento hierárquico dos dados de microarranjo é representado, frequentemente, como um mapa de calor – **Figura 2**.

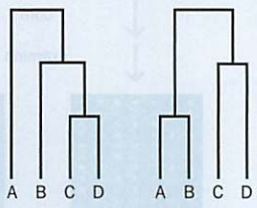

Figura 1 Análise dos dados de microarranjo utilizando os perfis de expressão hipotéticas de quatro genes A-D. Métodos de agrupamento hierárquico produzem diagramas com ramos (*dendogramas*) nos quais os genes com os perfis de expressão mais similares são agrupados em conjunto, mas métodos de agrupamento alternativos produzem dendogramas com diferentes topologias. O padrão da esquerda é típico da topologia produzida pelo agrupamento dos vizinhos mais próximos (ligação única); o padrão da direita é típico da topologia produzida pelo agrupamento dos vizinhos mais afastados (ligação completa).

Agrupamento não hierárquico

Uma desvantagem do agrupamento hierárquico é que ele exige muito tempo e muitos recursos. Como alternativa, os métodos não hierárquicos separam os dados de expressão em um número pré-definido de grupos. Como resultado, a análise é acelerada consideravelmente, especialmente quando o grupo de dados é muito grande. No *método de agrupamento k-médias*, vários pontos conhecidos como centros do grupo são definidos no início da análise, e cada gene é designado ao centro de grupo mais apropriado.

Com base em todos os membros de cada grupo, a média é recalculada (os centros dos grupos são reposicionados). A análise é então repetida até que todos os genes sejam designados a novos centros de grupo. Este processo é reiterado até que os membros de vários grupos não sofram mais mudanças. *Mapas de auto-organização* são similares em conceito, mas o algoritmo é refinado por meio do uso de uma rede neural.

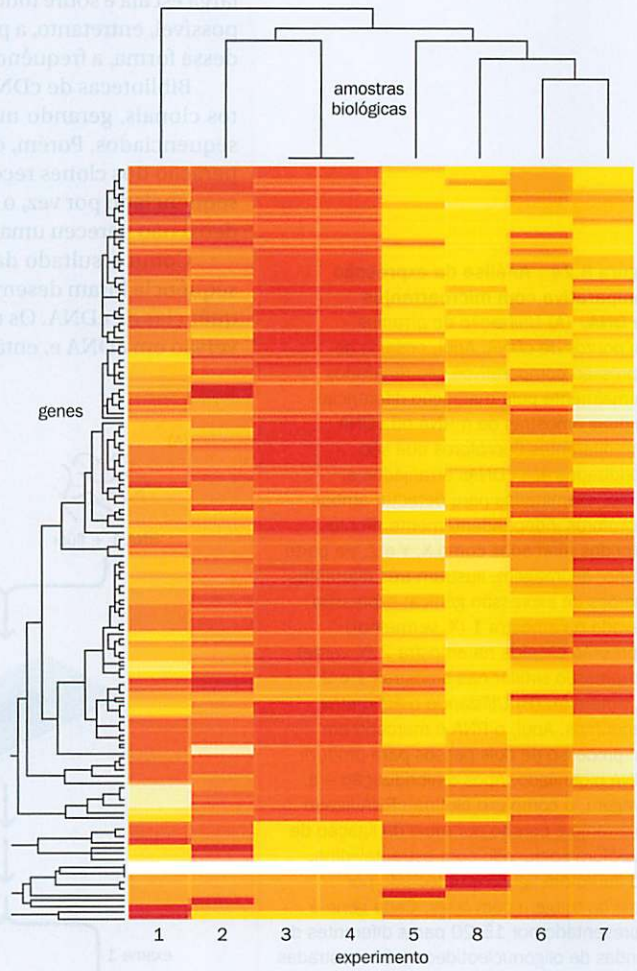

Figura 2 Mapas de calor como ferramenta para visualização da análise de microarranjo. *Mapas de calor* fornecem uma visão rápida dos grupos de genes que possuem valores de expressão similares. Eles consistem em pequenas células, cada uma com uma cor, as quais representam os valores de expressão relativos. Mapas de calor são frequentemente gerados a partir de análises de agrupamento hierárquico de diferentes amostras biológicas (geralmente dispostos em colunas, como mostrado aqui) e genes (normalmente em linhas e agrupados de acordo com a similaridade na expressão). [Adaptada de Allison DB, Cui X, Page GP & Sabripour M (2006) *Nat. Rev. Genet.* 7, 55-65. Com permissão de Macmillan Publishers Ltd. Ver a mesma referência para modos alternativos de visualização da análise de microarranjo.]

Há uma grande tendência de as sondas hibridizarem com outras sequências, além das sequências-alvo esperadas, sendo, portanto, necessária a utilização de controles adicionais. Da mesma forma, no sistema Affymetrix GeneChip, cada gene é representado por 20 ou mais diferentes sondas oligonucleotídicas selecionadas de diferentes regiões ao longo da sequência transcrita. Além dos 20 oligonucleotídeos de combinação perfeita (CP) por gene, uma série de 20 oligonucleotídeos correspondentes com combinação imperfeita (CI) é desenhada, mudando-se uma única base de cada sequência CP para ser utilizada como controle da hibridização inespecífica (**Figura 8.24B**). Para determinar o sinal para um gene em particular, o sinal de todos os 20 oligonucleotídeos CP é adicionado em conjunto e os sinais de todos os 20 oligonucleotídeos CI são subtraídos do total.

A caracterização moderna de expressão gênica global utiliza progressivamente o sequenciamento para quantificar transcritos

A hibridização por microarranjo forneceu uma fantástica nova tecnologia que, no final de 1990, foi a primeira a oferecer a habilidade de conduzir a caracterização de transcritos em larga escala e sobre todo o genoma. A caracterização global da expressão gênica também é possível, entretanto, a partir do sequenciamento direto de cópias de cDNA de transcritos; dessa forma, a frequência de transcritos pode ser quantificada para genes individuais.

Bibliotecas de cDNA têm sido caracterizadas pelo sequenciamento parcial de insertos clonais, gerando numerosas sequências EST, ou, algumas vezes, insertos totais são sequenciados. Porém, devido ao trabalho envolvido na geração de bibliotecas e na recuperação dos clones recombinantes, além do pequeno número de amostras que pode ser sequenciado por vez, o sequenciamento padrão dos clones da biblioteca pelo método dideoxi não pareceu uma maneira interessante de quantificar transcritos.

Como resultado das dificuldades mencionadas, vários métodos de amostragem de sequência foram desenvolvidos para reduzir os esforços envolvidos na recuperação de sequências de cDNA. Os métodos se baseiam na obtenção de RNA de fontes celulares, conversão em cDNA e, então, recuperação de pequenas sequências (*sequências marcadoras*)

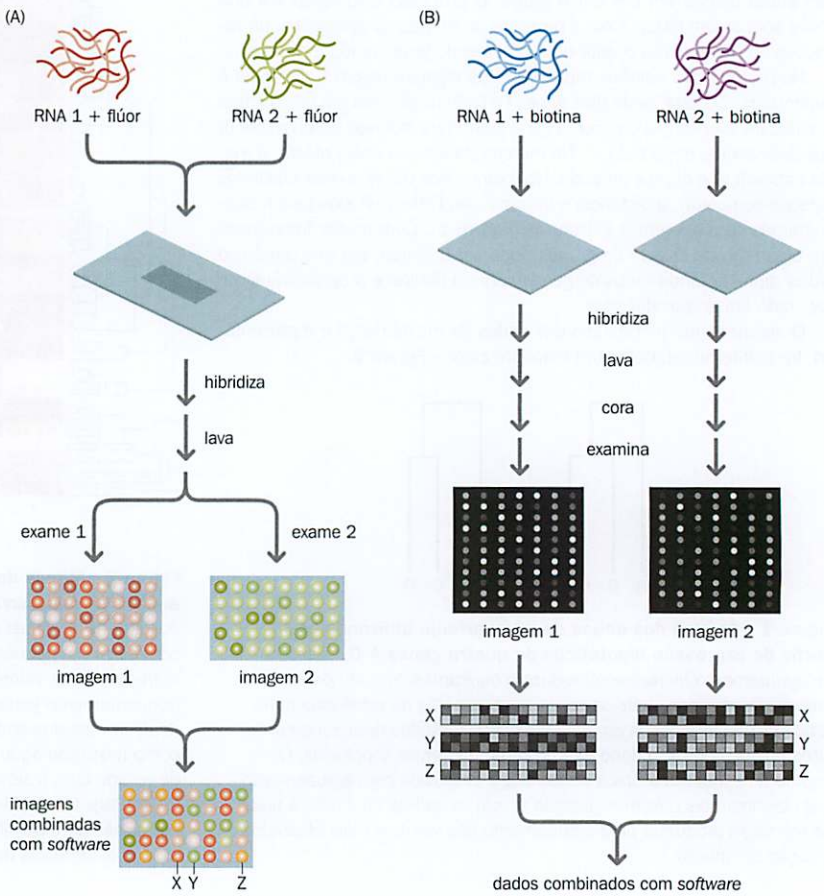

Figura 8.24 Análise da expressão comparativa com microarranjos de DNA. (A) *Utilização de arranjos em pontos de cDNA*. Aqui, ensaios de expressão comparativos são realizados normalmente pela marcação diferencial de duas amostras de mRNA ou cDNA com diferentes fluoróforos que são hibridizados aos cDNAs arranjados e, então, examinados para detectar ambos fluoróforos independentemente. Pontos coloridos marcados como X, Y e Z, na parte inferior da imagem, ilustram três diferentes padrões de expressão gênica: expressão elevada na amostra 1 (X, vermelho), expressão elevada na amostra 2 (Y, verde) e expressão similar nas amostras 1 e 2 (Z, amarelo). (B) *Utilizando o Affymetrix GeneChips*. Aqui, o RNA é marcado em um processo de dois passos para produzir cRNA biotinilado. Após a hibridização e a lavagem, o complexo biotina-cRNA ligado ao arranjo é corado por meio da ligação de fluoróforo conjugado com estreptavidina, e o fluoróforo ligado é detectado por meio da triagem com *laser*. Cada gene é representado por 15-20 pares diferentes de sondas de oligonucleotídeos (são mostradas 16 neste exemplo). Um membro de cada par é uma sonda de oligonucleotídeo com complementaridade perfeita ao gene; o outro é um oligonucleotídeo controle com uma modificação proposital. O exemplo mostra dados de expressão para três genes hipotéticos, representando genes que são preferencialmente expressos na amostra 1 (X), preferencialmente expressos na amostra 2 (Y) ou que demonstram expressão equivalente nas amostras 1 e 2 (Z). [Adaptada de Harrington CA, Rosenow C & Retief J (2000) *Curr. Opin. Microbiol.* 3, 285-291. Com permissão de Elsevier.]

de cDNAs individuais que sejam representativos de transcritos e cujas frequências sejam uma medida quantitativa da expressão daquele transcrito.

Um método engenhoso de amostragem de sequência popular é conhecido como análise em série da expressão gênica (SAGE, do inglês *serial analysis of gene expression*). Nesta técnica, as sequências marcadoras são obtidas a partir de várias cópias de cDNA e ligadas em longas séries para formar concatâmeros, os quais são então sequenciados. No método alternativo, o sequenciamento massivo de assinaturas em paralelo (MPSS, do inglês *massively parallel signature sequencing*), um grande número de cDNAs são ligados a microesferas em um fluxo celular e o sequenciamento de pequenas sequências marcadoras é realizado em paralelo, em vez de em série, como no SAGE.

O advento recente do sequenciamento massivo de DNA em paralelo significa que tanto a caracterização de expressão baseada em microarranjo como o sequenciamento de amostras em série (SAGE) serão extintos em um futuro próximo. A habilidade de sequenciar centenas de milhões de amostras em paralelo oferece um grande poder na quantificação de transcritos. Como não há a necessidade de amplificação de sequência por PCR, o sequenciamento individual de moléculas pode, em última análise, fornecer a quantificação mais acurada de transcritos.

A expressão proteica global é geralmente caracterizada com eletroforese em gel bidimensional e espectrometria de massa

A análise do transcriptoma é, claramente, bastante útil para acompanhar a expressão gênica, mas os produtos finais de muitos genes são proteínas. Níveis proteicos não refletem facilmente os níveis de transcritos de um gene codificante de proteína, pois há diferenças nas taxas de *turnover* proteico entre diferentes transcritos codificantes de proteínas. A atividade proteica também depende do nível de transcrito correspondente.

Assim como a transcriptômica, a proteômica pode ser utilizada para monitorar a abundância de diferentes produtos gênicos. A expressão de todas as proteínas celulares pode ser comparada entre amostras relacionadas, permitindo que padrões similares de expressão proteica possam ser identificados, evidenciando importantes mudanças que ocorrerem no proteoma, por exemplo, durante uma doença ou a resposta a um estímulo externo em particular. Isto é chamado, algumas vezes, de expressão proteômica.

Ao contrário do genoma, o proteoma varia amplamente entre diferentes tipos celulares em um organismo. Células humanas contêm dezenas de milhares de proteínas que diferem em abundância em quatro ou mais ordens de magnitude. Assim como os ácidos nucleicos, proteínas podem ser detectadas e identificadas por interações moleculares específicas, em muitos casos com a utilização de anticorpos ou outros ligantes como sondas. Entretanto, ao contrário dos ácidos nucleicos, não há um procedimento para clonagem e amplificação de proteínas raras. Além disso, propriedades físicas e químicas de proteínas são tão diversas que nenhuma metodologia única e universal análoga à hibridização pode ser utilizada para estudar o proteoma inteiro em um único experimento.

Atualmente, a expressão proteômica é amplamente baseada na tecnologia de "separação e revelação": misturas proteicas complexas são separadas em seus componentes para que características interessantes (como proteínas presentes em uma amostra doente, mas ausentes em uma amostra saudável correspondente) possam ser selecionadas para posterior caracterização. A separação é realizada, normalmente, com eletroforese em gel bidimensional, uma técnica bem padronizada que tem o poder de determinar até 10 mil proteínas em um único gel. Subsequentemente, as proteínas de interesse separadas podem ser submetidas à espectrometria de massa.

Eletroforese em gel de poliacrilamida bidimensional (2D-PAGE)

O princípio do 2D-PAGE é separar proteínas de acordo com a sua carga (primeira dimensão) e depois de acordo com a sua massa (segunda dimensão, perpendicularmente em relação à primeira). Uma amostra proteica complexa é carregada em um gel de poliacrilamida desnaturante e separada na primeira dimensão de acordo com o foco isoelétrico. Nesta técnica, as proteínas migram em um gradiente de pH até alcançarem o seu ponto isoelétrico, a posição onde a sua carga é neutra em relação ao pH local.

O procedimento padrão é preparar um gel com gradiente de pH imobilizado, onde os grupos de tamponamento são acoplados à matriz de poliacrilamida (para prevenir que eles fiquem à deriva e se tornem estáveis em corridas de gel mais longas). O gel é então equilibrado no detergente SDS, o qual é utilizado para quebrar ligações não covalentes de proteínas para que todas elas assumam conformações alongadas similares; ligando-se estequio-

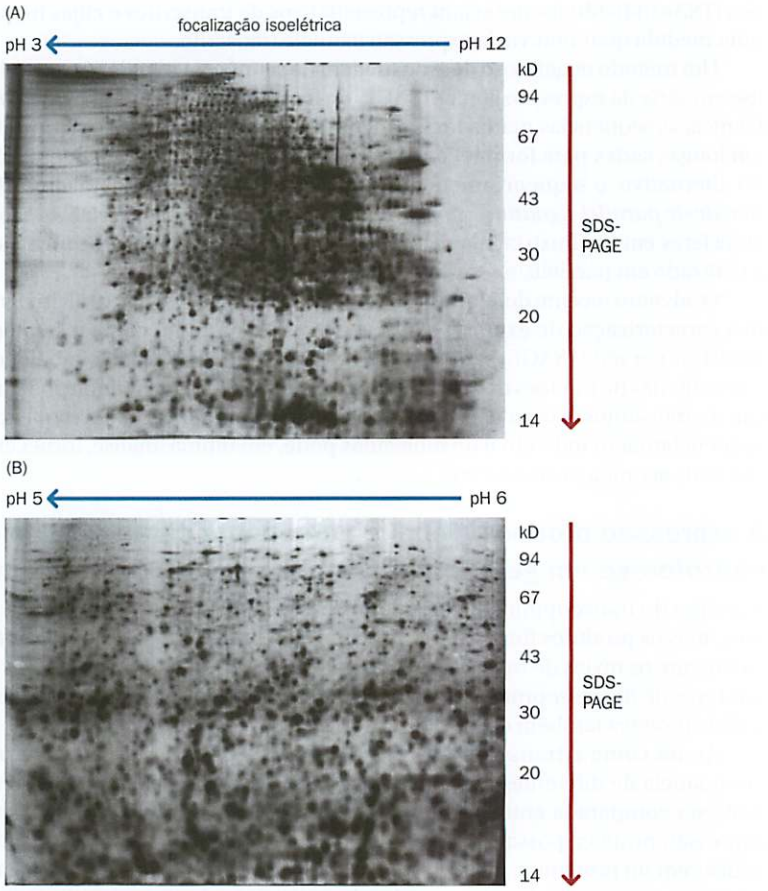

Figura 8.25 Eletroforese em gel bidimensional. Na eletroforese em gel bidimensional as amostras proteicas são, inicialmente, fracionadas por focalização isoelétrica (em um tubo com gel, ou em uma tira pré-moldada, compreendendo uma matriz de poliacrilamida com um gradiente de pH imobilizado). As frações separadas são então submetidas à eletroforese em gel de poliacrilamida com SDS em uma segunda dimensão, perpendicular à primeira. Ambas as imagens representam proteínas de fígado de camundongo separadas por eletroforese em gel bidimensional e coradas com prata para revelar pontos proteicos individuais. (A) Em um gel com um intervalo de pH grande (pH 3-12), as proteínas são agrupadas no centro, pois a maioria das proteínas possuem pontos isoelétricos no intervalo de pH de 4-7. (B) Um gel com um intervalo de pH estreito (pH 5-6) possui maior poder de resolução. [De Orengo CA, Jones DT & Thornton JM et al. (2003) Bioinformatics: Genes, Proteins And Computers. BIOS Scientific Publishers.]

metricamente ao *backbone* das proteínas desnaturadas, o SDS confere uma carga massiva negativa que cancela, efetivamente, quaisquer diferenças de cargas entre proteínas. A separação na segunda dimensão é, portanto, dependente da massa da proteína, onde proteínas menores se movem mais prontamente através dos poros do gel. O gel é então corado e as proteínas são reveladas como um complexo padrão de pontos (**Figura 8.25**).

Embora o 2D-PAGE tenha uma alta resolução e seja a técnica mais amplamente utilizada para separação de proteínas na proteômica, há várias limitações à sua utilidade em termos de representação, sensibilidade, reprodutibilidade e conveniência. Várias classes de proteínas estão sub-representadas em géis padrão, inclusive proteínas bastante básicas, proteínas com pouca solubilidade em tampões aquosos e proteínas de membrana.

A sensibilidade do 2D-PAGE é dependente do limite de detecção para proteínas muito escassas, mas a classe de corantes SYPRO® pode detectar pontos proteicos na faixa de nanogramas. A sensibilidade também é influenciada pela resolução do gel, pois os pontos que representam proteínas escassas podem ser obscurecidos por aqueles que representam proteínas em abundância. O pré-fracionamento pode remover proteínas em abundância, simplificando, assim, a amostra inicial carregada no gel, e a resolução pode ser melhorada a partir da utilização de géis com uma faixa bastante estreita de pH (ver Figura 8.25).

Desenvolvimentos importantes para melhorar a eficiência do 2D-PAGE incluem *softwares* para reconhecimento e quantificação de pontos, algoritmos que podem comparar pontos proteicos entre múltiplos géis e robôs que podem escolher pontos de interesse e processá-los para posterior identificação por espectrometria de massa, como descrito a seguir. Entretanto, a principal limitação do 2D-PAGE é que ele não é altamente adequado para automatização, tornando difícil a realização de análises de muitas amostras com alta eficiência. Métodos de separação alternativos baseados em cromatografia líquida de alto desempenho (HPLC, do inglês *multidimensional high-performance liquid chromatography*) podem, por fim, substituir o 2D-PAGE como a principal plataforma para separação proteica.

Espectrometria de massa

A espectrometria de massa (MS, do inglês *Mass Spectrometry*) é utilizada para determinar a massa exata de moléculas em uma amostra em particular. Na expressão proteômica ela

ajuda a identificar proteínas em pontos selecionados que tenham sido resolvidos pelo 2D-PAGE. Isso é feito pela determinação acurada da massa molecular, um procedimento que pode ser realizado mais rapidamente do que o sequenciamento direto de proteínas pelo método de degradação de Edman. Também é fácil de automatizar a MS para análise de amostras com alta eficiência, podendo assim processar, convenientemente, milhares de pontos de um gel bidimensional ou centenas de frações de HPLC.

Até recentemente, a MS não podia ser aplicada a grandes moléculas, como proteínas e ácidos nucleicos, pois estes eram quebrados em fragmentos aleatórios durante o processo de ionização. Essa limitação foi superada utilizando-se os métodos de ionização branda como a dessorção/ionização a *laser* assistida por matriz (MALDI, do inglês *Matrix-assisted Laser Desorption/Ionization*) (**Figura 8.26** e **Quadro 8.8**) e a ionização por *electrospray* (ESI, do inglês *Electrospray Ionization*).

As massas das proteínas ou, mais frequentemente, os fragmentos peptídicos derivados das proteínas por meio da digestão com proteases, podem ser utilizadas para identificar as proteínas, correlacionando-se massas determinadas experimentalmente com aquelas preditas em bancos de dados de sequências. Como descrito a seguir, há três maneiras diferentes de anotação de proteínas por EM (**Figura 8.27**):

- **Impressão digital de massa de peptídeos (PMF,** do inglês *Peptide Mass Fingerprinting*) é mais adequada para proteomas simples. Uma mistura simples de proteínas (um único ponto de um gel bidimensional) é digerida com tripsina para gerar uma coleção de peptídeos trípticos. Estes são submetidos ao MALDI-MS utilizando um analisador por tempo de voo (TOF, do inglês *Time-of-Flight*) (ver Figura 8.26 e Quadro 8.8), o qual retorna um conjunto de espectros de massa. Os espectros são utilizados como entrada na busca de proteínas em bancos de dados de sequências. O algoritmo de busca realiza digestões virtuais com tripsina de todas as proteínas do banco de dados e calcula as massas dos peptídeos trípticos preditos. Ele tenta, então, corresponder essas massas preditas com aquelas determinadas experimentalmente.
- A **busca por fragmento iônico** é mais adequada à análise de proteomas complexos, sendo que o algoritmo pode ser modificado para levar em conta as massas das modificações pós-traducionais conhecidas. Os fragmentos peptídicos trípticos são analisados pela espectrometria de massa em *tandem* (MS/MS; ver Quadro 8.8), durante a qual os peptídeos são quebrados em fragmentos aleatórios. Os espectros de massa desses fragmentos podem ser utilizados na busca contra bancos de dados de ESTs (os quais não podem ser verificados com os dados de PMF, pois os peptídeos intactos são geralmente muito grandes). Qualquer resultado de EST pode, então, ser utilizado em uma busca pelo BLAST para identificar a sequência inteira de possíveis homólogos. Um algoritmo específico chamado MS-BLAST é útil para lidar com as assinaturas de sequências curtas obtidas dos fragmentos iônicos de peptídeos.
- O **sequenciamento *de novo* de peptídeos-escada** também é realizado, pois é impossível levar em conta todas as variantes, seja a nível de sequência (p. ex., polimorfismos) como a nível de modificação proteica (p. ex., glicanas complexas). O sequenciamento de peptídeos-escada pode fornecer assinaturas de sequência que podem ser utilizadas como entrada na busca da identificação de proteínas homólogas em bancos de dados. Nesta técnica, os fragmentos peptídicos gerados por MS/MS são

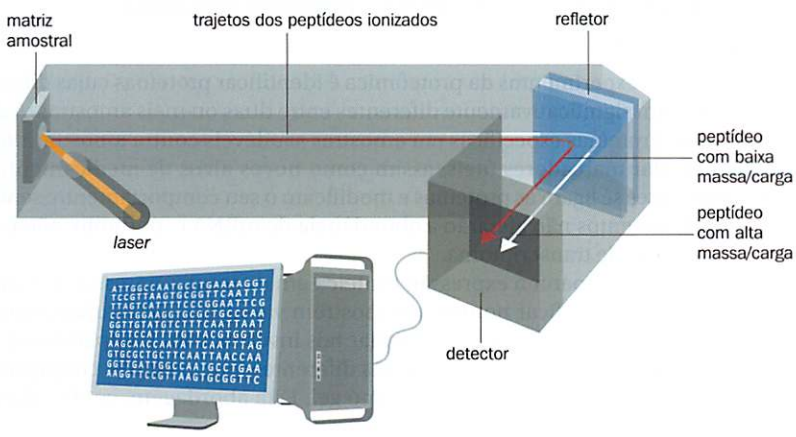

Figura 8.26 Princípio da espectroscopia de massa MALDI-TOF. O *analito* (a amostra sob estudo – normalmente uma coleção de fragmentos peptídicos trípticos) é misturado com um componente de matriz e colocado próximo à fonte de *laser*. O *laser* aquece o analito – cristais de matriz, expandindo-o na fase gasosa sem que ocorra fragmentação significativa. Os íons, então, percorrem um tubo de voo em direção a um refletor, o qual foca os íons em um detector. O tempo de voo (o tempo despendido para que os íons alcancem o detector) é dependente da taxa representada pela divisão da massa pela carga (massa/carga) e permite que a massa de cada molécula do analito seja gravada.

QUADRO 8.8 Espectrometria de massa na proteômica

O espectrômetro de massa

Um espectrômetro de massa possui três componentes. Um *ionizador* converte a amostra a ser analisada (o analito) em íons na fase gasosa, acelerando-os contra o *analisador de massa*. Este último separa os íons de acordo com a taxa representada pela divisão da massa pela carga (massa/carga) no caminho para o *detector de íon*, o qual grava o impacto de íons individuais, apresentando-os como um espectro de massa do analito.

A ionização de grandes moléculas sem a fragmentação e degradação é conhecida como *ionização branda*. Dois métodos de ionização branda são amplamente utilizados na proteômica.

- **Dessorção/ionização a *laser* assistida por matriz (MALDI)** envolve a mistura do analito (p. ex., peptídeos trípticos oriundos de uma amostra em particular) com uma matriz de absorção luminosa em um solvente orgânico. A evaporação do solvente produz cristais de analito/matriz, os quais são aquecidos por um pulso curto de energia a *laser*. A dessorção da energia a *laser* na forma de calor promove a expansão da matriz e do analito em fase gasosa. O analito é então ionizado e acelerado contra o detector.
- Na **ionização por *electrospray* (ESI)**, o analito é dissolvido, e a solução é empurrada em um capilar estreito. Uma diferença de potencial, aplicada através da abertura, permite que o analito saia como um *spray* fino de partículas carregadas. As gotas evaporam conforme os íons adentram no analisador de massa.

Analisadores de massa

Um **analisador de massa quadripolar** é composto por quatro cilindros metálicos, pares os quais são conectados eletricamente e carregam voltagens opostas que podem ser controladas pelo operador. Os espectros de massa são obtidos variando-se a diferença potencial aplicada por meio do feixe de íons, permitindo que íons com diferentes taxas massa/carga sejam direcionados para o detector. Um **analisador de tempo de voo (TOF)** mede o tempo que os íons levam para percorrer o trajeto ao longo do cilindro em direção ao detector, um fator que depende da taxa massa/carga.

Espectrometria de massa em *tandem* (MS/MS)

Dois ou mais analisadores de massa são operados em série. Vários instrumentos MS/MS têm sido descritos, incluindo instrumentos de quadripolos triplos e híbridos quadripolo/tempo de voo. Os analisadores de massa são separados por uma célula de colisão que contém gás inerte, promovendo a dissociação de íons em fragmentos. O primeiro analisador seleciona um íon de um peptídeo em particular e o direciona para dentro de uma célula de colisão, onde ele é fragmentado. Um espectro de massa para os fragmentos é então obtido pelo segundo analisador. Essas duas funções podem ser combinadas em instrumentos mais sofisticados, como os analisadores ciclotrônicos de captura de íons (*ion-trap*) de Transformada de Fourier.

arranjados em grupos aninhados e diferem em tamanho por um único aminoácido. A partir da comparação das massas desses fragmentos com tabelas padrão de aminoácidos, é possível deduzir a sequência do fragmento peptídico *de novo*, mesmo quando uma sequência com combinação precisa não é encontrada no banco de dados. Na prática, a abordagem do sequenciamento *de novo* é complicada pela presença de duas séries de fragmentos, uma aninhada na extremidade N-terminal e uma na extremidade C-terminal. As duas séries podem ser distinguidas acoplando-se marcadores de massa diagnósticos a uma das extremidades da proteína.

Análises comparativas de expressão proteica possuem muitas aplicações

O objetivo de muitos experimentos da proteômica é identificar proteínas cujas diferenças na abundância sejam significativamente diferentes entre duas ou mais amostras. A abundância variável de proteínas específicas em amostras saudáveis contra amostras doentes pode ajudar a revelar marcadores úteis, assim como novos alvos de medicamentos. Os fármacos muitas vezes se ligam às proteínas e modificam o seu comportamento; sendo assim, muitos medicamentos não afetarão a abundância de mRNA e, portanto, não podem ser analisados ao nível de transcriptoma.

Uma maneira de comparar a expressão proteica em diferentes amostras é examinar géis bidimensionais e identificar pontos que mostrem variação quantitativa (vários pacotes de *softwares* estão disponíveis para auxiliar tais investigações comparativas). Uma alternativa é marcar proteínas de duas amostras diferentes, por exemplo, conjugando-as com Cy3 e Cy5 e, então, separando-as no mesmo gel. Esta abordagem, conhecida como

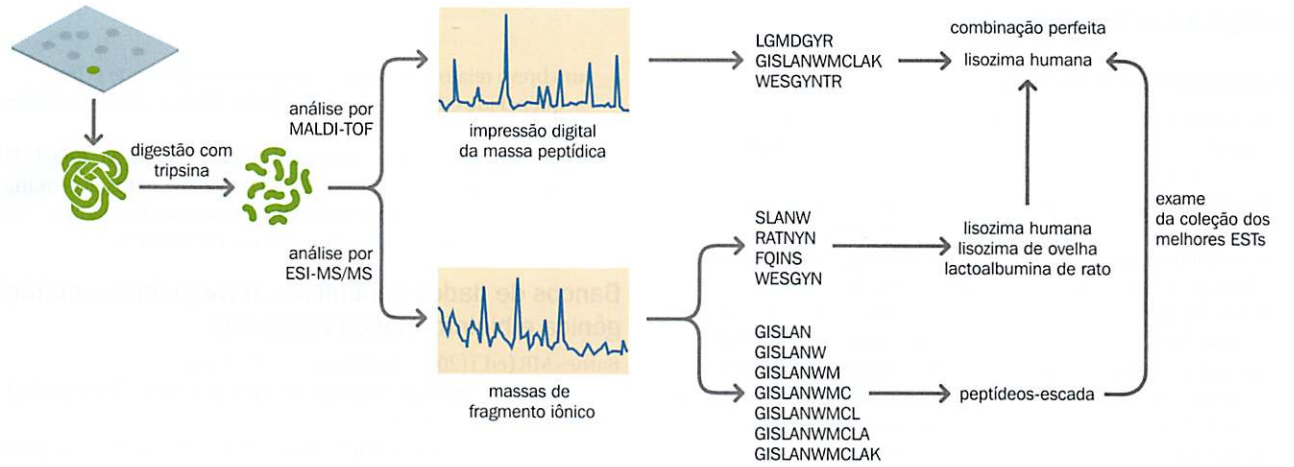

Figura 8.27 Anotação de proteínas por espectroscopia de massa. Amostras proteicas individuais (p.ex., pontos de géis bidimensionais) são digeridas com tripsina, a qual cliva a extremidade C-terminal de resíduos de lisina (K) ou arginina (R), contanto que o próximo resíduo não seja prolina. Os peptídeos trípticos podem ser analisados como moléculas intactas pelo método MALDI-TOF, e as massas utilizadas como entradas na busca em bancos de dados de proteínas. São utilizados algoritmos que pegam a sequência proteica, clivando-as com a mesma especificidade da tripsina, e comparam as massas teóricas desses peptídeos com as massas experimentais obtidas por MS. Idealmente, as massas de vários peptídeos devem identificar a mesma proteína parente que, nesse caso, é a lisozima humana. Pode não haver melhores resultados se a proteína não estiver no banco de dados ou, mais provavelmente, se tiver sido submetida a modificações pós-traducionais ou modificações artificiais durante o experimento. Nessas circunstâncias, a ionização com *electrospray* acoplada com espectroscopia de massa em *tandem* (ESI-MS/MS) pode ser utilizada para fragmentar os íons. As massas dos fragmentos iônicos podem ser utilizadas como entradas na busca em bancos de dados de ESTs e obtenção de combinações parciais, as quais podem levar, por fim, à correta anotação. Alternativamente, as massas de peptídeos-escada podem ser utilizadas para determinar as sequências proteicas *de novo*.

eletroforese em gel diferencial (DIGE, do inglês *Difference Gel Electrophoresis*), explora o mesmo princípio da expressão gênica diferencial com microarranjos de DNA.

Outra abordagem é marcar proteínas de diferentes fontes com marcadores de afinidade dependentes de isótopos (ICATs, do inglês I*sotope-Coded Affinity Tags*). Um espectrômetro de massa pode distinguir e quantificar facilmente duas formas marcadas com isótopos do mesmo componente que são quimicamente idênticas e podem ser copurificadas. Um método ICAT foi desenvolvido, portanto, utilizando-se um derivado de iodoacetamida biotinilada para marcar misturas proteicas de modo seletivo nos resíduos de cisteína. A partir da ligação com estreptavidina, a cauda de biotina permite a purificação por afinidade dos peptídeos marcados com cisteína após a proteólise com tripsina.

O reagente de ICAT está disponível nas formas marcadas com isótopos pesados e leves, os quais podem ser utilizados para marcar *pools* celulares diferenciais sob diferentes condições (como saudável *versus* doente). Após a marcação, as células são reunidas e lisadas, e as proteínas, isoladas, assim perdas ocorridas na purificação são equivalentes para ambas as amostras. Se elas são equivalentes é porque não ocorreu superexpressão ou baixa expressão e, assim, a proteína não é de interesse imediato. Se as intensidades diferirem, uma mudança na expressão proteica aconteceu, e a proteína é de interesse. A quantidade das duas formas é mensurada e a forma peptídica leve é fragmentada e identificada a partir da busca em banco de dados.

A primeira década de 2000 será vista como uma época revolucionária para a análise do genoma. A tecnologia tem avançado para permitir o rápido mapeamento e sequenciamento de genomas complexos, sendo que sequências de genomas inteiros foram obtidas para vários genomas complexos, inclusive para o genoma humano. O advento dos métodos de sequenciamento massivo em paralelo de DNA significa que o sequenciamento genômico será, comparativamente, rápido no futuro; talvez em torno de 2015-2020 o sequenciamento individual de genoma será rotineiro e barato. Além disso, análises em larga escala começaram a identificar todos os genes em genomas e catalogar todos os seus produtos de expressão, dando nome a diferentes transcritos de RNA e proteínas traduzidas. Cada tipo de produto de expressão pode ser detectado dentro de células intactas (por hibridização ou imunocitoquímica) e quantificado por técnicas como PCR em tempo real, *western blot* ou espectrometria de massa. Refinamentos atuais têm permitido a análise de números enormes de amostras, sendo que agora um número enorme de genes é estudado em paralelo em um único experimento. Independentemente do número de genes estudados e da metodologia empregada, o principal objetivo permanece o mesmo: entender como os genes atuam e como a doença altera a produção celular de transcritos e proteínas.

LEITURAS ADICIONAIS

Sequenciamento do DNA

Clarke J, Wu HC, Jayasinghe L et al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270.

Eid J, Fehr A, Gray J et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.

Gupta P (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602–611.

Hodges E, Xuan Z, Balija V et al. (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527.

Mamanova L, Coffey AJ, Scott CE et al. (2010) Target-enrichment strategies for next generation sequencing. *Nat. Meth.* 7, 111–118.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141.

Metzker M (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.* 11, 31–46.

Tucker T, Marra M, Friedman JM (2009) Massively parallel DNA sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* 85, 142–154.

Projeto Genoma Humano e mapas genéticos humanos (ver também Quadro 8.2 para referências históricas adicionais)

All About The Human Genome Project (HGP). http://www.genome.gov/10001772 [Um recurso pedagógico mantido pelo *US National Human Genome Research Institute.*]

Garber M, Zody MC, Arachchi HM et al. (2009) Closing gaps in the human genome using sequencing by synthesis. *Genome Biol.* 10, R60.

Genome Reference Consortium. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/ [Fornece atualizações e dados gerais em gráficos a respeito dos projetos de genoma humano e de camundongos.]

Imanishi T, Itoh T, Suzuki Y et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* 2, e162.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861.

Diversos autores (2001) Human Genome Issue. *Nature* 409, 813–958. [Disponível em Nature Network OmicsGateway portal, http://www.nature.com/omics/subjects/genomesequenceandanalysis/archive/2001/index.html]

Diversos autores (2001) Human Genome Issue. *Science* 291, 1177–1351 [Artigos disponíveis em http://www.sciencemag.org/content/vol291/issue5507/index.dtl]

Diversos autores (2003) User's Guide to the Human Genome. *Nature Genet.* 35 (1s), 1–79. [Disponível em http://www.nature.com/ng/journal/v35/n1s/index.html]

Diversos autores (2006) Human Genome Collection. *Nature* S1, 1–305. [Contém artigos que analisam a sequência de todos os 24 cromossomos humanos, disponível em http://www.nature.com/nature/supplements/collections/humangenome/index.html]

Waterston RH, Lander ES & Sulston JE (2002) On the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* 99, 3712–3716.

Projetos Genoma de Organismos-modelo

Genome News Network. Um guia rápido referente a genomas já sequenciados. http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_index.shtml. [Fornece um breve relato sobre cada organismo-modelo que já foi sequenciado e *links* com a bibliografia relativa a esses modelos e genomas.]

Liolios K, Mavrommatis K, Tavernarakis N & Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 36 (Database issue), D475–D479.

Bancos de dados genômicos, navegadores, anotação gênica e bioinformática associada

Barnes MR (ed.) (2007) Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data, 2nd ed. John Wiley and Sons.

Bina M (2006) Use of genome browsers to locate your favorite genes. *Methods Mol. Biol.* 338, 1–7.

Reeves GA, Talavera D & Thornton JM (2009) Genome and proteome annotation: organization, interpretation and integration. *J. R. Soc. Interface* 6, 129–147.

Spudich G, Fernández-Suárez XM & Birney E (2007) Genome browsing with Ensembl: a practical overview. *Brief. Funct. Genomics Proteomics* 6, 202–219.

The Gene Ontology Consortium (2008) The Gene Ontology Project in 2008. *Nucleic Acids Res.* 36, D440–D444. [Ver também http://www.geneontology.org/]

Wheeler DL, Barrett T, Benson DA et al. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36 (Database issue), D13–D21.

Análises básicas da expressão gênica

Applied Biosystems (undated) Real-Time PCR Vs. Traditional PCR. http://www.appliedbiosystems.com/support/tutorials/pdf/rtpcr_vs_tradpcr.pdf [Um tutorial para a PCR em tempo real.]

VanGuilder HD, Vrana KE & Freeman WM (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques* 44, 619–626.

Ward TH & Lippincott-Schwartz J (2006) The uses of green fluorescent protein in mammalian cells. *Methods Biochem. Anal.* 47, 305–337.

Análises avançadas da expressão gênica em paralelo: caracterização de transcritos

Allison DB, Cui X, Page GP & Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7, 55–65.

Belacel N, Wang Q & Cuperlovic-Culf M (2006) Clustering methods for microarray gene expression data. *Omics* 10, 507–531.

Jongeneel CV, Delorenzi M, Iseli C et al. (2005) An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res.* 15, 1007–1014.

Murray D, Doran P, MacMathuna P & Moss AC (2007) *In silico* gene expression analysis—an overview. *Mol. Cancer* 6, 50.

Various authors (2002) The Chipping Forecast II. *Nat. Genet.* 32 (Suppl.), 465–551.

Análises avançadas da expressão gênica em paralelo: caracterização de proteínas

Hamdan M & Righetti PG (2003) Assessment of protein expression by means of 2-D gel electrophoresis with and without mass spectrometry. *Mass Spectrom. Rev.* 22, 272–284.

Kolker E, Higdon R & Hogan JM (2006) Protein identification and expression analysis using mass spectrometry. *Trends Microbiol.* 14, 229–235.