CAPÍTULO 3

O Genoma Humano: Estrutura e Função Gênicas

Ao longo das últimas 3 décadas, houve um progresso marcante na compreensão da estrutura e função dos genes e cromossomos. Esses avanços foram auxiliados pelas aplicações da genética molecular e da genômica em muitos problemas clínicos, fornecendo as ferramentas para uma nova abordagem distinta para a genética médica. Neste capítulo, apresentamos uma visão geral da estrutura e função gênicas e dos aspectos da genética molecular que são necessários para a compreensão das abordagens genéticas e genômicas na medicina. Para complementar as informações discutidas aqui e nos capítulos subsequentes, fornecemos material *on-line* adicional para detalhar muitas das abordagens experimentais da genética e genômica modernas que estão se tornando críticas para a prática e compreensão da genética humana e médica.

O maior conhecimento dos genes e da sua organização no genoma teve um impacto enorme na medicina e na nossa percepção da fisiologia humana. Em 1980, Paul Erb foi agraciado com o prêmio Nobel por ter previsto o início desta nova era:

> *Como o nosso conhecimento e nossa prática atuais da medicina dependem de um conhecimento sofisticado de anatomia, fisiologia e bioquímica humanas, lidar com a doença no futuro exigirá uma compreensão detalhada da anatomia, fisiologia e bioquímica moleculares do genoma humano... Necessitaremos de um conhecimento mais detalhado de como os genes humanos são organizados e como funcionam e são regulados. Teremos também de ter médicos que estejam tão familiarizados com a anatomia molecular e fisiologia dos cromossomos e genes como o cirurgião cardíaco está familiarizado com a estrutura e funcionamento do coração.*

INFORMAÇÕES DO CONTEÚDO DO GENOMA HUMANO

Como o código digital de três bilhões de letras do genoma humano orienta os detalhes da anatomia, fisiologia e bioquímica humanas, às quais Berg se refere? A resposta está nas enormes amplificação e integração do conteúdo de informações que ocorre quando se passa dos genes no genoma para os seus produtos na célula e para a expressão observável dessa informação genética, como traços celulares, morfológicos, clínicos ou bioquímicos — o que é denominado **fenótipo** do indivíduo. Essa expansão hierárquica de informações do genoma para o fenótipo inclui uma vasta gama de produtos de RNA estruturais e reguladores, bem como produtos proteicos que orquestram as muitas funções das células, órgãos e todo o organismo, além de suas interações com o meio ambiente. Mesmo com a sequência essencialmente completa do genoma humano em mãos, ainda não sabemos o número exato de genes no genoma. As estimativas atuais são de que o genoma contenha cerca de 20.000 **genes codificadores de proteínas** (veja o Quadro no Cap. 2), mas esse retrato começa somente a sugerir os níveis de complexidade que emergem da decodificação dessa informação digital (Fig. 3-1).

Como introduzido brevemente no Capítulo 2, o produto de genes codificadores de proteínas é uma proteína, cuja estrutura por fim determina as suas funções específicas na célula. Mas se houvesse uma simples correspondência de um para um entre genes e proteínas, poderíamos ter no máximo cerca de 20.000 proteínas diferentes. Esse número parece insuficiente para dar conta da vasta gama de funções que ocorre em células humanas ao longo da vida. A resposta para esse dilema é encontrada em duas características da estrutura e função gênicas. Em primeiro lugar, muitos genes são capazes de gerar vários produtos diferentes, não apenas um (Fig. 3-1). Esse processo, discutido mais adiante neste capítulo, é efetuado através do uso de segmentos de codificação alternativos nos genes e de modificações bioquímicas subsequentes da proteína codificada; essas duas características dos genomas complexos resultam em uma amplificação substancial do conteúdo de informações. Na verdade, estima-se que, dessa maneira, os 20.000 genes humanos podem codificar muitas centenas de milhares de proteínas diferentes, coletivamente chamadas de **proteoma**. Em segundo lugar, proteínas individuais não funcionam sozinhas. Elas formam redes elaboradas, envolvendo muitas proteínas diferentes e RNAs reguladores que respondem de maneira coordenada e integrada a muitos diferentes sinais genéticos, ambientais ou de desenvolvimento. A natureza combinatória das redes de proteínas resulta em uma diversidade ainda maior de possíveis funções celulares.

Os genes estão localizados ao longo do genoma, mas tendem a se agrupar em regiões e em cromossomos específicos e a ser relativamente escassos em outras regiões

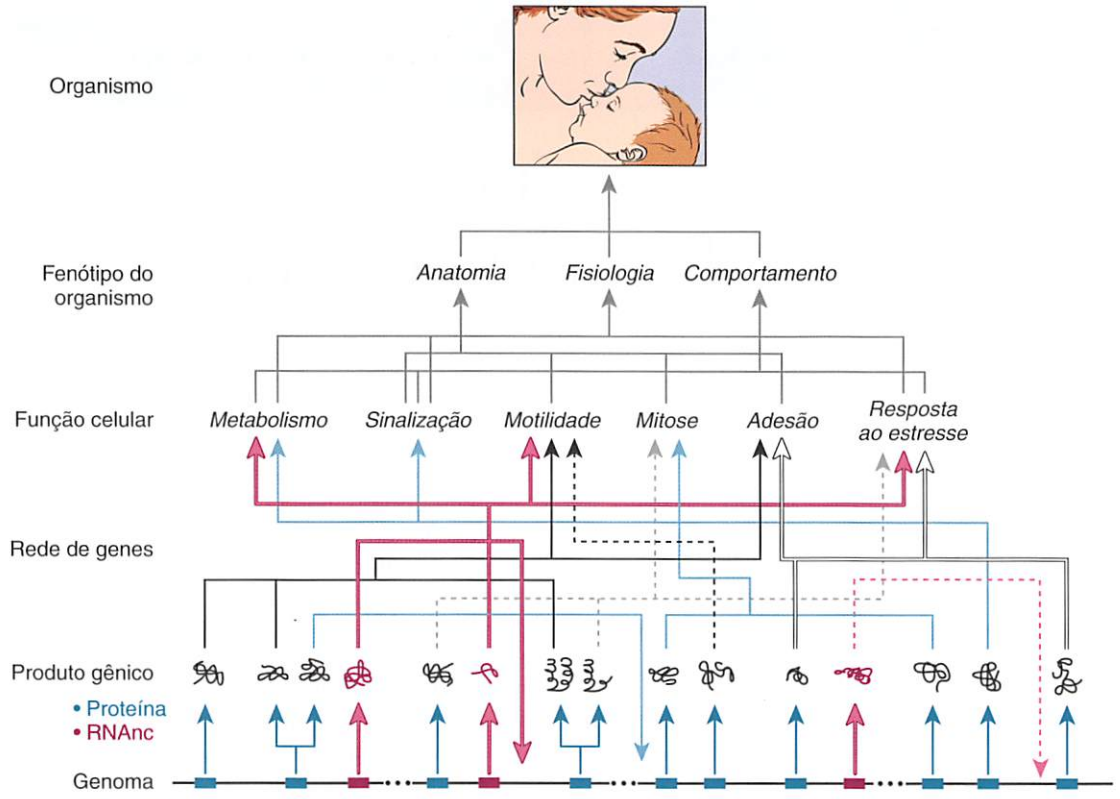

Figura 3-1 **Amplificação da informação genética do genoma para os produtos gênicos, para as redes de genes e, finalmente, para a função celular e fenótipo.** O genoma contém tanto genes de RNA codificantes de proteínas (*em azul*) como genes de RNA não codificantes (RNAnc) (*em vermelho*). Muitos genes no genoma usam informações de codificação alternativas para gerar vários produtos diferentes. RNAncs grandes e pequenos participam da regulação gênica. Muitas proteínas participam em redes multigênicas que respondem aos sinais celulares de maneira coordenada e combinatória, ampliando ainda mais a gama de funções celulares subjacentes aos fenótipos do organismo.

ou em outros cromossomos. Por exemplo, o cromossomo 11, que possui aproximadamente 135 milhões de pb (pares de megabase [Mb]), é relativamente rico em genes, com cerca de 1.300 genes que codificam proteínas (Fig. 2-7). Esses genes não estão distribuídos aleatoriamente ao longo do cromossomo, e sua localização é particularmente aumentada em duas regiões cromossômicas com densidade gênica tão alta quanto um gene a cada 10 kb (Fig. 3-2). Alguns desses genes pertencem a famílias de genes relacionados, como descreveremos com mais detalhes posteriormente neste capítulo. Outras regiões são pobres em genes e existem vários dos chamados desertos de genes, de um milhão de pares de bases ou mais, sem qualquer gene codificante de proteína conhecido. Duas advertências aqui: em primeiro lugar, o processo de identificação do gene e a anotação do genoma ainda são um desafio contínuo; apesar da aparente robustez de estimativas recentes, é praticamente certo que existem alguns genes, incluindo genes clinicamente relevantes, que atualmente não são detectados ou que apresentam características que atualmente não são reconhecidas como sendo associadas a genes. E, em segundo lugar, como mencionado no Capítulo 2, muitos genes não são codificantes de proteínas; seus produtos são moléculas de RNA funcionais (**RNAs não codificadores** ou RNAnc; Fig. 3-1), que desempenham uma variedade de funções na célula, muitas das quais estão apenas começando a ser desvendadas.

Para genes localizados nos autossomos, existem duas cópias de cada gene, uma no cromossomo herdado da mãe e uma no cromossomo herdado do pai. Para a maioria dos genes autossômicos, ambas as cópias são expressas e geram um produto. Existe, no entanto, um número crescente de genes no genoma que são exceções a essa regra geral e são expressos a partir das duas cópias em níveis caracteristicamente diferentes, incluindo alguns que, em caso extremo, são expressos a partir de apenas um dos dois homólogos. Esses exemplos de **desequilíbrio alélico** são discutidos detalhadamente adiante neste capítulo, bem como nos Capítulos 6 e 7.

O DOGMA CENTRAL: DNA → RNA → PROTEÍNA

Como o genoma especifica a complexidade e diversidade funcionais evidentes na Figura 3-1? Como vimos no capítulo anterior, a informação genética está contida no DNA nos

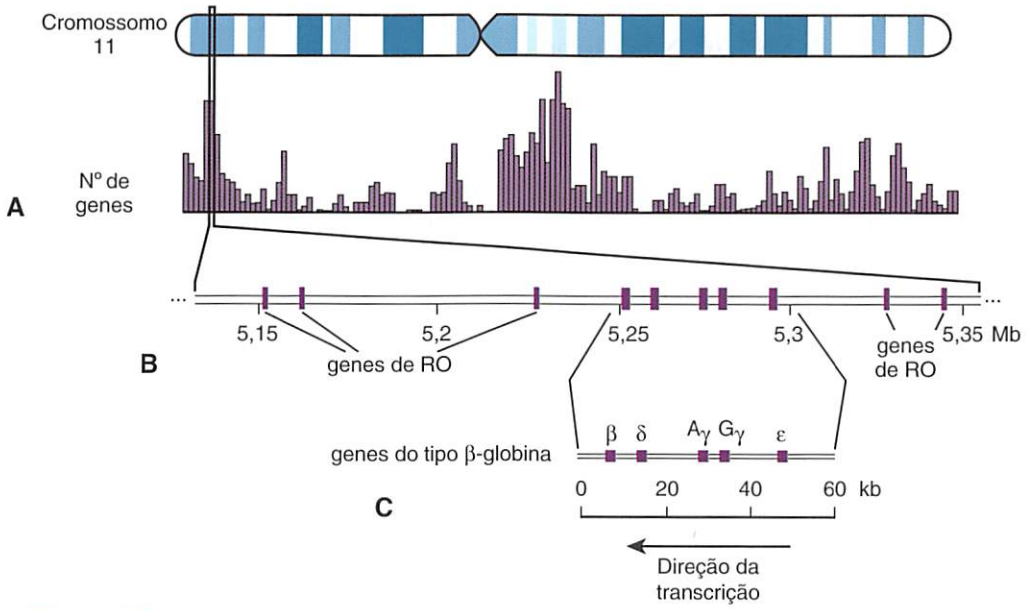

Figura 3-2 **Conteúdo gênico do cromossomo 11, que consiste em 135 Mb de DNA. A,** A distribuição dos genes é indicada ao longo do cromossomo e é alta em duas regiões do cromossomo e baixa nas demais regiões. **B,** Uma região expandida de 5,15 a 5,35 Mb (medida a partir do telômero do braço curto), que contém 10 genes codificantes de proteínas conhecidos, cinco pertencentes à família gênica do receptor olfativo (RO) e cinco pertencentes à família gênica da globina. **C,** Os cinco genes do tipo β-globina expandiram-se ainda mais. *Veja Fontes & Agradecimentos.*

cromossomos dentro do núcleo celular. No entanto, a síntese proteica, o processo pelo qual a informação codificada no genoma é efetivamente utilizada para especificar funções celulares, ocorre no citoplasma. Essa compartimentalização reflete o fato de que o organismo humano é um **eucarionte**. Isto significa que as células humanas possuem um núcleo que contém o genoma, separado do citoplasma por uma membrana nuclear. Ao contrário, nos procariontes, como a bactéria intestinal *Escherichia coli*, o DNA não está inserido dentro de um núcleo. Devido à compartimentalização de células eucarióticas, a transferência de informações do núcleo para o citoplasma é um processo complexo que tem sido foco de muita atenção entre biólogos moleculares e celulares.

A ligação molecular entre esses dois tipos relacionados de informação — o código do DNA dos genes e o código do aminoácido da proteína — é o **ácido ribonucleico (RNA)**. A estrutura química do RNA é semelhante à do DNA, exceto que cada nucleotídeo no RNA tem um componente de açúcar ribose no lugar de uma desoxirribose; além disso, a uracila (U) substitui a timina como uma das bases de pirimidina do RNA (Fig. 3-3). Outra diferença entre o RNA e o DNA é que o RNA, na maioria dos organismos, existe como uma molécula de fita única, enquanto o DNA, como vimos no Capítulo 2, existe como uma dupla-hélice.

As relações de informações entre o DNA, o RNA e as proteínas estão interligadas: o DNA genômico direciona a síntese e a sequência de RNA, o RNA direciona a síntese e sequência de polipeptídeos, e as proteínas específicas estão envolvidas na síntese e no metabolismo do DNA e do RNA. Esse fluxo de informações é chamado de **dogma central** da biologia molecular.

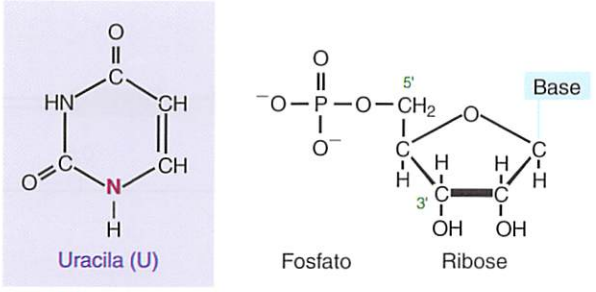

Figura 3-3 **A pirimidina uracila e a estrutura de um nucleotídeo no RNA.** Observe que o açúcar ribose substitui o açúcar desoxirribose do DNA. Compare com a Figura 2-2.

A informação genética está armazenada no DNA do genoma por meio de um código (o **código genético**, discutido adiante), no qual a sequência de bases adjacentes por fim determina a sequência de aminoácidos no polipeptídeo codificado. Primeiramente, o RNA é sintetizado a partir do molde de DNA por um processo conhecido como **transcrição**. O RNA, que carrega a informação codificada sob a forma chamada de **RNA mensageiro (RNAm)**, é então transportado do núcleo para o citoplasma, onde a sequência de RNA é decodificada, ou traduzida, para determinar a sequência de aminoácidos na proteína que está sendo sintetizada. O processo de **tradução** ocorre nos **ribossomos**, que são organelas citoplasmáticas com locais de ligação para todas as moléculas de interação, incluindo o RNAm, envolvido na síntese proteica. Os ribossomos são compostos de muitas proteínas estruturais diferentes em associação com tipos especializados de RNA

conhecidos como **RNA ribossômicos (RNAr)**. A tradução envolve ainda um terceiro tipo de RNA, o **RNA de transferência (RNAt)**, que fornece a ligação molecular entre o código contido na sequência de bases de cada RNAm e a sequência de aminoácidos da proteína codificada por tal RNAm.

Devido ao fluxo interdependente de informações representado pelo dogma central, pode-se começar a discussão da genética molecular da expressão gênica em qualquer um dos seus três níveis de informação: DNA, RNA ou proteína. Começamos examinando a estrutura dos genes no genoma como uma base para a discussão do código genético, transcrição e tradução.

ORGANIZAÇÃO E ESTRUTURA GÊNICAS

De forma mais simples, um gene codificante de proteína pode ser visualizado como um segmento de uma molécula de DNA que contém um código para uma sequência de aminoácidos de uma cadeia polipeptídica e as sequências reguladoras necessárias para a sua expressão. Essa descrição, no entanto, é inadequada para genes no genoma humano (e para a maioria dos genomas eucariontes), porque poucos genes existem como sequências codificantes contínuas. Em vez disso, na maioria dos genes, as sequências codificantes são interrompidas por uma ou mais regiões não codificantes (Fig. 3-4). Essas sequências interpostas, chamadas de

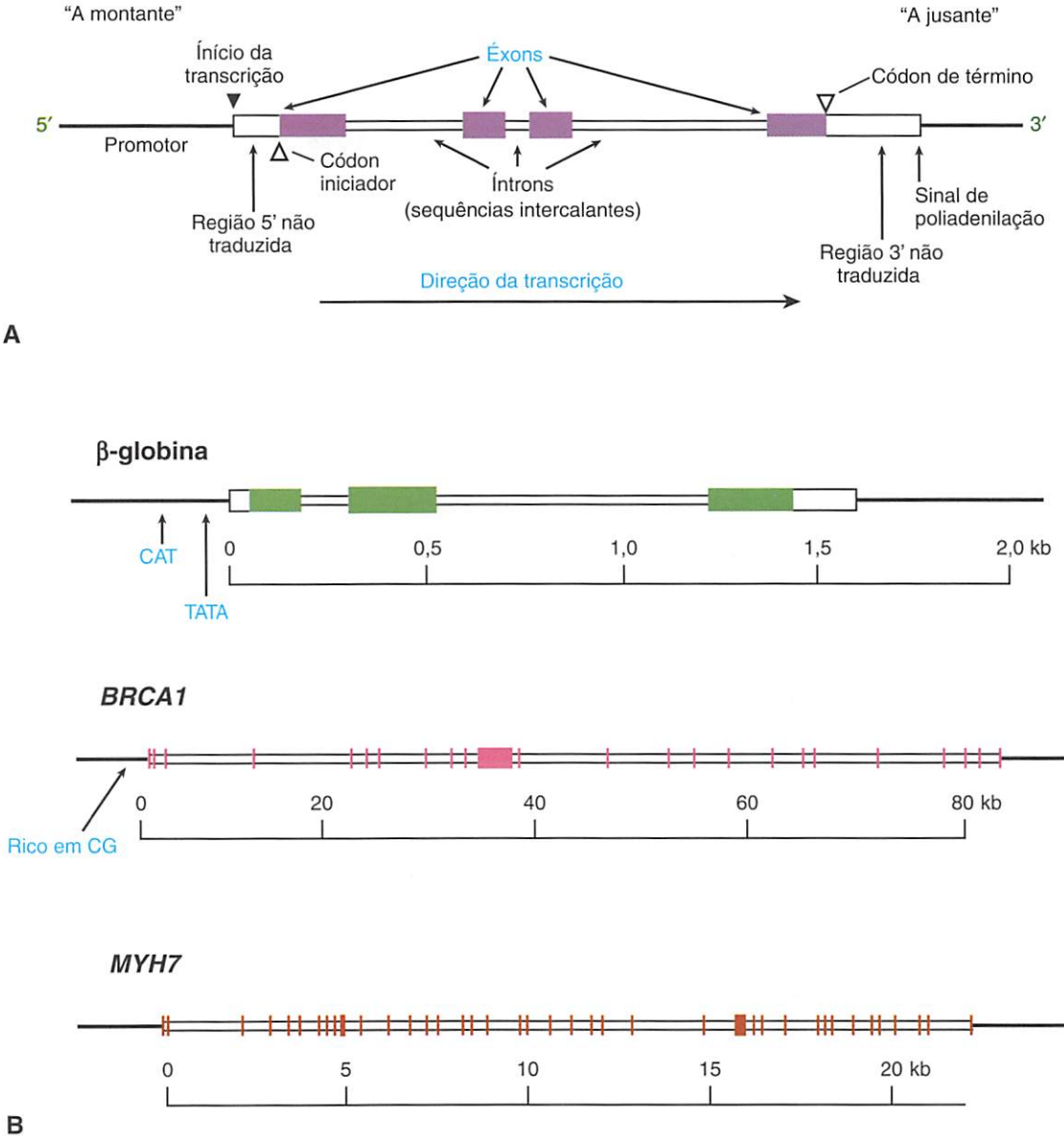

Figura 3-4 A, Estrutura geral de um gene humano típico. As características individuais marcadas são discutidas no texto. B, Exemplos de três genes humanos clinicamente importantes. Diferentes mutações no gene da β-globina, com três éxons, causam uma variedade de distúrbios importantes de hemoglobina (Casos 42 e 44). As mutações no gene *BRCA1* (24 éxons) são responsáveis por vários casos de câncer de mama e de ovário hereditários (Caso 7). As mutações no gene da cadeia pesada de β-miosina (*MYH7*) (40 éxons) levam à miocardiopatia hipertrófica hereditária.

íntrons, são inicialmente transcritas em RNA no núcleo, mas não estão presentes no RNAm maduro no citoplasma, porque são removidas ("spliced out") por um processo que discutiremos adiante. Assim, a informação de sequências intrônicas não é, normalmente, representada no produto final da proteína. Os íntrons são alternados com **éxons**, os segmentos de genes que determinam, por fim, a sequência de aminoácidos da proteína. Além disso, a coleção de éxons codificantes em qualquer gene em particular é flanqueada por sequências adicionais que são transcritas mas não traduzidas, chamadas de regiões não traduzidas 5′ e 3′ (Fig. 3-4). Embora alguns genes no genoma humano não tenham íntrons, a maioria dos genes contém pelo menos um e geralmente vários íntrons. Em muitos genes, o tamanho cumulativo dos íntrons compõe uma proporção muito maior do comprimento total de um gene do que os éxons. Embora alguns genes tenham apenas alguns pares de quilobases de tamanho, outros estendem-se por centenas de pares de quilobases. Além disso, alguns genes são excepcionalmente grandes; por exemplo, o gene da distrofina no cromossomo X (nos quais mutações levam à distrofia muscular de Duchenne [Caso 14]) abrange mais de 2 Mb, dos quais, notavelmente, menos de 1% consiste em éxons codificantes.

Características Estruturais de um Gene Humano Típico

Uma gama de aspectos caracteriza os genes humanos (Fig. 3-4). Nos Capítulos 1 e 2, definimos brevemente *gene* em termos gerais. Nesse momento, podemos fornecer uma definição molecular de um gene como uma *sequência de DNA que especifica a produção de um produto funcional*, seja um polipeptídeo ou uma molécula de RNA funcional. Um gene inclui não apenas as sequências codificantes de nucleotídeos reais, mas também as sequências de nucleotídeos adjacentes necessárias para a expressão adequada do gene, isto é, para a produção de RNAm normal ou de outras moléculas de RNA na quantidade correta, no local correto e no tempo correto durante o desenvolvimento ou durante o ciclo celular.

As sequências de nucleotídeos adjacentes fornecem os sinais moleculares de "início" e "parada" para a síntese de RNAm transcrito a partir do gene. Pelo fato de o transcrito de RNA primário ser sintetizado na direção de 5′ para 3′, o início da transcrição é chamado de extremidade 5′ da porção transcrita de um gene (Fig. 3-4). Por convenção, o DNA genômico que antecede o local de início de transcrição na direção 5′ é chamado de sequência "a montante" (*upstream*), enquanto que a sequência de DNA localizada na direção 3′ além da extremidade de um gene é chamada de sequência "a jusante" (*downstream*). Na extremidade 5′ de cada gene encontra-se uma região **promotora** que inclui sequências responsáveis pelo início adequado da transcrição. Dentro dessa região estão vários elementos de DNA, cuja sequência é frequentemente conservada entre vários genes diferentes; esta conservação, em conjunto com estudos funcionais de expressão gênica, indica que essas sequências específicas desempenham um papel importante na regulação gênica. Apenas um subconjunto de genes no genoma é expresso em qualquer tecido ou em qualquer momento durante o desenvolvimento. Vários tipos diferentes de promotor são encontrados no genoma humano, com diferentes propriedades reguladoras que especificam os padrões, bem como os níveis de expressão de um gene determinado em diferentes tecidos e tipos celulares, tanto durante o desenvolvimento como ao longo da vida. Algumas dessas propriedades são codificadas no genoma, enquanto outras são especificadas por características da cromatina associadas a essas sequências, conforme discutido mais adiante neste capítulo. Tanto os promotores quantos outros **elementos reguladores** (localizados tanto em 5′ ou 3′ de um gene ou em seus íntrons) podem ser locais de mutação em doenças genéticas que podem interferir na expressão normal de um gene. Esses elementos reguladores, incluindo os **acentuadores**, **os insuladores** e **as regiões de controle do *locus***, são discutidos detalhadamente mais adiante neste capítulo. Alguns desses elementos encontram-se a uma distância significativa da porção codificante de um gene, o que reforça o conceito de que o ambiente genômico no qual um gene está inserido é uma característica importante da sua evolução e regulação.

A região não traduzida 3′ contém um sinal para a adição de uma sequência de resíduos de adenosina (a chamada cauda poliA) à extremidade do RNA maduro. Embora geralmente seja aceito que essas sequências reguladoras estreitamente contíguas façam parte do que é chamado de gene, as dimensões precisas de qualquer gene em particular permanecerão um tanto incertas, até que as funções potenciais das sequências mais distantes sejam completamente caracterizadas.

Famílias de Genes

Muitos genes pertencem a famílias gênicas, que compartilham sequências de DNA estreitamente relacionadas e codificam polipeptídeos com sequências de aminoácidos estreitamente relacionadas.

Membros de duas dessas famílias gênicas estão localizados dentro de uma pequena região no cromossomo 11 (Fig. 3-2) e ilustram uma série de aspectos que caracteriza as famílias gênicas em geral. Uma família gênica pequena e clinicamente importante é composta de genes que codificam as cadeias de proteínas encontradas nas hemoglobinas. Acredita-se que o *cluster* (aglomerado) de genes da β-globina no cromossomo 11 e o aglomerado de genes relacionados da α-globina no cromossomo 16 tenham surgido pela duplicação de um gene precursor primitivo há cerca de 500 milhões de anos. Esses dois aglomerados contêm múltiplos genes que codificam cadeias de globina estreitamente relacionadas expressas em diferentes estágios do desenvolvimento, do embrião ao adulto. Acredita-se que cada aglomerado tenha evoluído por uma série de eventos sequenciais de duplicação gênica nos últimos 100 milhões de anos. Os padrões éxon-íntron dos genes funcionais de globina foram notavelmente conservados durante a evolução; cada um dos genes funcionais de globina possui dois íntrons em localizações semelhantes (veja o gene de β-globina na Fig. 3-4), embora as sequências contidas nos íntrons tenham acumulado muito

mais alterações de bases de nucleotídeos ao longo do tempo do que as sequências codificantes de cada gene. O controle da expressão dos vários genes de globina, no estado normal, bem como em muitos distúrbios hereditários da hemoglobina, é considerado em mais detalhes mais adiante neste capítulo e no Capítulo 11.

A segunda família gênica mostrada na Figura 3-2 é a família de genes de receptores olfativos (RO). Estima-se que existam até 1.000 genes de RO no genoma. Os RO são responsáveis pelo nosso sentido olfativo aguçado que pode reconhecer e distinguir milhares de substâncias químicas estruturalmente diversas. Os genes de RO são encontrados em todo o genoma em quase todos os cromossomos, embora mais da metade seja encontrada no cromossomo 11, incluindo uma série de membros da família próximos do aglomerado de β-globina.

Pseudogenes

Dentro tanto da família gênica de β-globina quanto de RO há sequências que são relacionadas com a globina funcional e genes de RO, mas que não produzem qualquer RNA funcional ou produto proteico. Sequências de DNA que se assemelham muito a genes conhecidos, mas não são funcionais, são chamadas de **pseudogenes,** e existem dezenas de milhares de pseudogenes relacionados com muitos genes e famílias gênicas diferentes localizados ao longo do genoma. Os pseudogenes são de dois tipos gerais, processados e não processados. Acredita-se que os **pseudogenes não processados** sejam subprodutos da evolução, representando genes "mortos" que antes eram funcionais, mas que agora são vestigiais, tendo sido inativados por mutações sequências codificantes ou reguladoras críticas. Ao contrário dos pseudogenes não processados, os **pseudogenes processados** são pseudogenes que foram formados, não por mutação, mas por um processo chamado de **retrotransposição,** que envolve a transcrição, a geração de uma cópia de DNA a partir do RNAm (o chamado **DNAc**) por transcrição reversa e, por fim, a integração dessas cópias de DNA no genoma em um local geralmente bastante distante do gene original. Como esses pseudogenes são criados por retrotransposição de uma cópia de DNA do RNAm processado, eles não possuem íntrons e não estão necessária ou geralmente no mesmo cromossomo (ou região cromossômica) como seu gene progenitor. Em muitas famílias gênicas, existem tantos ou mais pseudogenes quanto membros de genes funcionais.

Genes de RNA não Codificante

Como discutido anteriormente, muitos genes codificam proteínas e são transcritos nos RNAms que, por fim, são traduzidos em suas respectivas proteínas; seus produtos compreendem as enzimas, proteínas estruturais, receptores e proteínas reguladoras que são encontrados em vários tecidos e tipos celulares humanos. No entanto, tal como apresentado brevemente no Capítulo 2, existem outros genes, cujo produto funcional parece ser o próprio RNA (Fig. 3-1). Estes chamados **RNA não codificantes (RNAnc)** têm uma gama de funções nas células, embora muitos não tenham uma função identificada. Pelas estimativas atuais, existem cerca de 20.000 a 25.000 genes de RNAnc, além dos aproximadamente 20.000 genes codificantes de proteínas que foram introduzidos anteriormente. Assim, a coleção de RNAnc representa aproximadamente metade de todos os genes humanos identificados. O cromossomo 11, por exemplo, apresenta uma estimativa de ter 1.000 genes RNAnc, além de seus 1.300 genes codificantes de proteínas.

Alguns dos tipos de RNAnc desempenham papéis amplamente genéricos na infraestrutura celular, incluindo os RNAt e RNAr envolvidos na tradução de RNAm nos ribossomos, outros RNAs envolvidos no controle do *splicing* de RNA, e os pequenos RNAs nucleolares (RNApno) envolvidos na modificação de RNAr. Outros RNnc podem ser bastante longos (às vezes chamados de RNAnc longos ou **RNAInc**) e desempenham papéis na regulação gênica, no silenciamento gênico e em doenças humanas, como exploraremos com mais detalhes mais adiante neste capítulo.

Uma classe específica de pequenos RNAs de importância crescente são os **microRNAs (miRNA)**, RNAnc de apenas cerca de 22 bases de comprimento que suprimem a tradução de genes-alvo por meio da ligação a seus respectivos RNAms e regulam a produção de proteínas a partir do(s) transcrito(s)-alvo(s). Bem mais de 1.000 genes de miRNA foram identificados no genoma humano; alguns são evolutivamente conservados, ao passo que outros parecem ter origem bastante recente durante a evolução. Alguns

RNAS NÃO CODIFICANTES E DOENÇAS

A importância de vários tipos de RNAnc para a medicina é ressaltada por seus papéis em uma gama de doenças humanas, desde síndromes precoces do desenvolvimento até distúrbios que se manifestam na idade adulta.

- A deleção de um agrupamento de genes de miRNA no cromossomo 13 leva a uma forma de **síndrome de Feingold,** uma síndrome de desenvolvimento de defeitos esqueléticos e de crescimento, incluindo microcefalia, baixa estatura e anomalias digitais.
- As mutações no gene de miRNA *MIR96*, na região do gene crítica para a especificidade de reconhecimento de seu(s) RNAm(s)-alvo, pode resultar em **perda auditiva progressiva** em adultos.
- Níveis alterados de determinadas classes de miRNAs foram relatados em uma ampla variedade de cânceres, distúrbios do sistema nervoso central e doença cardiovascular (Cap. 15).
- A deleção de agrupamentos de genes de RNApno no cromossomo 15 resulta na **síndrome de Prader-Willi,** um distúrbio caracterizado por obesidade, hipogonadismo e comprometimento cognitivo (Cap. 6).
- A expressão anormal de um RNAInc no cromossomo 12 tem sido relatada em pacientes com uma doença associada à gravidez, chamada de **síndrome HELLP.**
- Deleção, expressão anormal e/ou alterações estruturais em diferentes RNAIncs com papéis na regulação da expressão gênica de longo alcance e função genômica levam a uma variedade de distúrbios que envolvem a manutenção do tamanho do telômero, a expressão monoalélica de genes em regiões específicas do genoma e a dosagem do cromossomo X (Cap. 6).

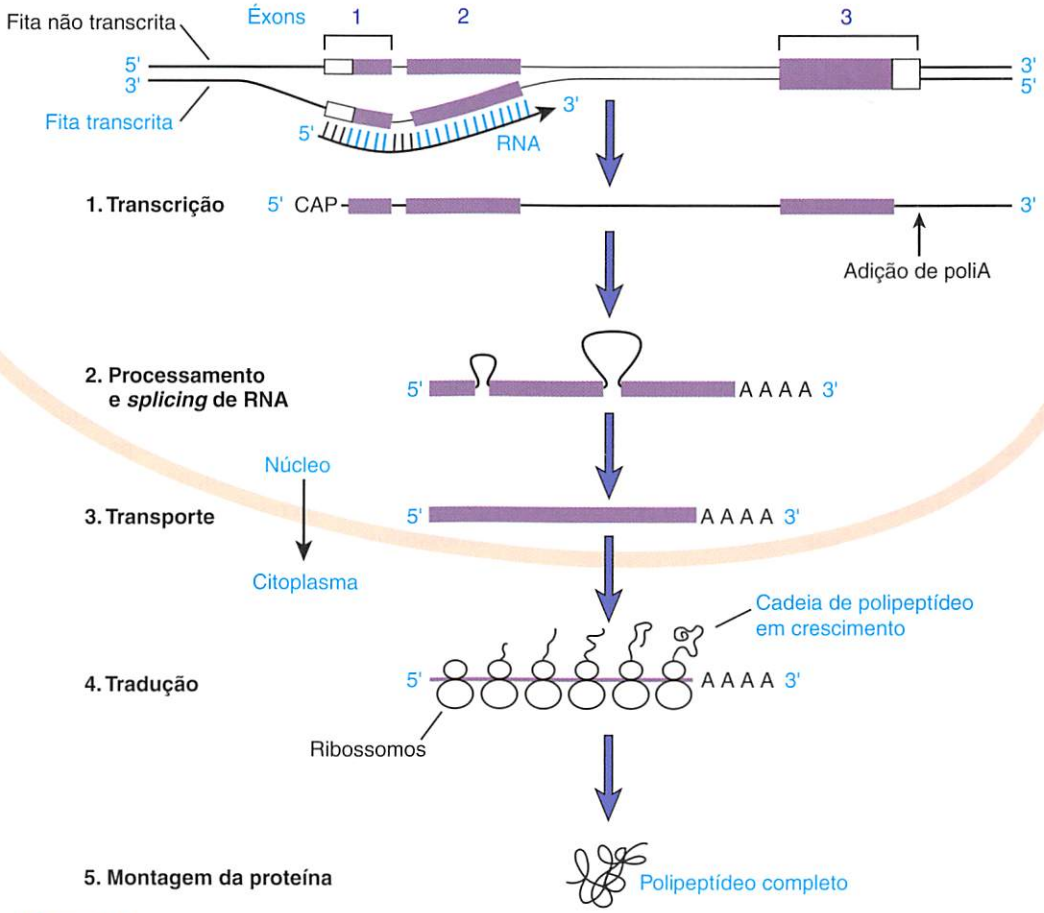

Figura 3-5 **Fluxo de informação do DNA até o RNA e até a proteína para um gene hipotético com três éxons e dois íntrons.** Dentro dos éxons, a cor *roxa* indica as sequências codificantes. As etapas incluem transcrição, processamento e *splicing* de RNA, transporte de RNA do núcleo para o citoplasma, e tradução.

miRNAs mostraram regular negativamente centenas de RNAms cada, com diferentes combinações de RNAs-alvo em diferentes tecidos; combinados, prevê-se que os miRNAs, portanto, controlem a atividade de até 30% de todos os genes codificantes de proteínas no genoma.

Embora esta seja uma área em rápido movimento da biologia genômica, mutações em vários genes de RNAnc já foram implicadas em doenças humanas, incluindo câncer, distúrbios do desenvolvimento e várias doenças tanto de início precoce como no adulto (Quadro).

FUNDAMENTOS DA EXPRESSÃO GÊNICA

Para genes que codificam proteínas, o fluxo de informações do gene para o polipeptídeo envolve vários passos (Fig. 3-5). O início da transcrição de um gene está sob a influência de promotores e outros elementos reguladores, bem como de proteínas específicas conhecidas como **fatores de transcrição**, que interagem com sequências específicas dentro dessas regiões e determinam um padrão espacial e temporal de expressão de um gene. A transcrição de um gene é iniciada no sítio de "início" de transcrição no DNA cromossômico no início de uma região 5′ transcrita, mas não traduzida (chamada de 5′ UTR), imediatamente a montante das sequências codificantes. Ela continua ao longo do cromossomo para qualquer lugar das várias centenas de pares base até mais de um milhão de pares de bases, passando tanto por íntrons como éxons, além da extremidade das sequências codificantes. Após a modificação nas extremidades 5′ e 3′ do transcrito de RNA primário, as porções correspondentes aos íntrons são removidas e os segmentos correspondentes aos éxons são removidos em conjunto, um processo chamado de ***splicing* de RNA**. Após o *splicing*, o RNAm resultante (contendo um segmento central que é agora colinear com as porções codificantes do gene) é transportado do núcleo para o citoplasma, onde o RNAm é finalmente traduzido em uma sequência de aminoácidos do polipeptídeo codificado. Cada uma das etapas dessa via complexa está sujeita a erros, e mutações que interferem nas etapas individuais têm sido implicadas em vários distúrbios hereditários (Caps. 11 e 12).

Transcrição

A transcrição de genes codificantes de proteínas pela RNA polimerase II (uma das várias classes de RNA polimerases) é iniciada no sítio de início transcricional, o ponto na 5′ UTR que corresponde à extremidade 5′ do produto final

de RNA (Figs. 3-4 e 3-5). A síntese do transcrito de RNA primário prossegue na direção de 5′ para 3′, enquanto a fita do gene que é transcrita e que serve como molde para a síntese de RNA é na verdade lida na direção de 3′ a 5′ em relação à direção do arcabouço de desoxirribose fosfodiéster (Fig. 2-3). Como o RNA sintetizado corresponde tanto em polaridade quanto em sequência de bases (substituindo T por U) à fita 5′ a 3′ do DNA, esta fita de 5′ a 3′ de DNA não transcrito é às vezes chamada de fita de DNA *codificante*, ou **senso**. A fita de DNA de 3′ a 5′ que é usada como molde para a transcrição é então chamada de fita *não codificante* ou **antissenso**. A transcrição continua por ambas as porções intrônicas e exônicas do gene, para além da posição no cromossomo que, por fim, corresponde à extremidade 3′ do RNAm maduro. Não se sabe se a transcrição termina em um ponto de término 3′ predeterminado.

O transcrito primário de RNA é processado pela adição de uma estrutura química de "cap" (ou capuz) na extremidade 5′ do RNA e pela clivagem da extremidade 3′ em um ponto específico a jusante da extremidade da informação de codificação. Essa clivagem é seguida pela adição de uma cauda poliA à extremidade 3′ do RNA; a cauda poliA parece aumentar a estabilidade do RNA poliadenilado resultante. A localização do ponto de poliadenilação é especificada em parte pela sequência AAUAAA (ou uma variante desta), geralmente encontrada na porção 3′ não traduzida do transcrito de RNA. Todas essas modificações pós-transcricionais ocorrem no núcleo, assim como o processo de *splicing* de RNA. O RNA totalmente processado, chamado agora de RNAm, é então transportado para o citoplasma, onde ocorre a tradução (Fig. 3-5).

Tradução e Código Genético

No citoplasma, o RNAm é traduzido em uma proteína pela ação de uma variedade de pequenas moléculas adaptadoras de RNA, os RNAts, cada qual específico para um aminoácido em particular. Essas moléculas notáveis, cujo tamanho varia de apenas 70 a 100 nucleotídeos, têm a tarefa de trazer os aminoácidos corretos para a posição correta ao longo do molde de RNAm, para serem adicionados à cadeia polipeptídica em crescimento. A síntese proteica ocorre nos ribossomos, complexos macromoleculares compostos de RNAr (codificados pelos genes de RNAr 18S e 28S) e várias dúzias de proteínas ribossômicas (Fig. 3-5).

A chave para a tradução é um código que relaciona aminoácidos específicos com combinações de três bases adjacentes ao longo do RNAm. Cada conjunto de três bases constitui um **códon**, específico para um determinado aminoácido (Tabela 3-1). Teoricamente, variações quase

TABELA 3-1 O Código Genético

	Segunda Base								
Primeira Base	U		C		A		G		Terceira Base
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC	phe	UCC	ser	UAC	tyr	UGC	cys	C
	UUA	leu	UCA	ser	UAA	stop	UGA	Stop	A
	UUG	leu	UCG	ser	UAG	stop	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC	leu	CCC	pro	CAC	his	CGC	arg	C
	CUA	leu	CCA	pro	CAA	gln	CGA	arg	A
	CUG	leu	CCG	pro	CAG	gln	CGG	arg	G
	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC	ile	ACC	thr	AAC	asn	AGC	ser	C
	AUA	ile	ACA	thr	AAA	lys	AGA	arg	A
	AUG	met	ACG	thr	AAG	lys	AGG	arg	G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC	val	GCC	ala	GAC	asp	GGC	gly	C
	GUA	val	GCA	ala	GAA	glu	GGA	gly	A
	GUG	val	GCG	ala	GAG	glu	GGG	gly	G

Abreviaturas para Aminoácidos			
ala (A)	alanina	leu (L)	leucina
arg (R)	arginina	lys (K)	lisina
asn (N)	asparagina	met (M)	metionina
asp (D)	ácido aspártico	phe (F)	fenilalanina
cys (C)	cisteína	pro (P)	prolina
gln (Q)	glutamina	ser (S)	serina
glu (E)	ácido glutâmico	thr (T)	treonina
his (H)	glicina	trp (W)	triptofano
gly (G)	histidina	tyr (Y)	tirosina
ile (I)	isoleucina	val (V)	valina

Stop, códon de término.
Os códons são apresentados em termos de RNAm, que são complementares aos códons de DNA correspondentes.

infinitas são possíveis no arranjo das bases ao longo de uma cadeia de polinucleotídeos. Em qualquer posição, existem quatro possibilidades (A, T, C ou G); assim, para três bases, existem 4^3, ou 64, possíveis combinações de trincas. Esses 64 códons constituem o **código genético.**

Como existem apenas 20 aminoácidos e 64 códons possíveis, a maioria dos aminoácidos é especificada por mais de um códon; portanto, o código é considerado **degenerado.** Por exemplo, a base na terceira posição da trinca frequentemente pode ser uma purina (A ou G) ou uma pirimidina (T ou C) ou, em alguns casos, qualquer uma das quatro bases, sem alterar a mensagem codificada (Tabela 3-1). A leucina e a arginina são, cada uma, especificadas por seis códons. Apenas a metionina e o triptofano são, cada um, especificados por um único códon. Três dos códons são chamados de **códons de parada** (ou ***nonsense***) porque designam o término da tradução do RNAm naquele ponto.

A tradução de um RNAm processado é sempre iniciada em um códon que especifica metionina. A metionina é, portanto, o primeiro aminoácido codificado (aminoterminal) de cada cadeia polipeptídica, embora seja geralmente removida antes de a síntese de proteínas ser concluída. O códon para metionina (o **códon iniciador,** AUG) estabelece a **matriz de leitura** do RNAm; cada códon subsequente é lido na sua vez para predizer a sequência de aminoácidos da proteína.

Os elos moleculares entre códons e aminoácidos são as moléculas de RNAt específicas. Um local determinado em cada RNAt forma um **anticódon** de três bases que é complementar a um códon específico no RNAm. A ligação entre o códon e o anticódon leva o aminoácido adequado à próxima posição no ribossomo para a fixação, pela formação de uma ligação peptídica na extremidade carboxílica da cadeia polipeptídica crescente. O ribossomo, em seguida, desliza exatamente três bases ao longo do RNAm, alinhando o próximo códon para reconhecimento por outro RNAt contendo o próximo aminoácido. Assim, proteínas são sintetizadas da extremidade aminoterminal até a extremidade carboxiterminal, o que corresponde à tradução do RNAm na direção 5′ a 3′.

Conforme mencionado anteriormente, a tradução termina quando um códon de parada (UGA, UAA ou UAG) é encontrado na mesma matriz de leitura que o códon iniciador. (Códons de parada em qualquer uma das outras matrizes de leitura não utilizadas não são lidos e, portanto, não têm efeito sobre a tradução.) O polipeptídeo completo é então liberado do ribossomo, que se torna disponível para iniciar a síntese de outra proteína.

Transcrição do Genoma Mitocondrial

As seções anteriores descreveram fundamentos da expressão gênica para genes contidos no genoma nuclear. O genoma mitocondrial possui transcrição e sistema de síntese de proteínas próprios. Uma RNA polimerase especializada, codificada no genoma nuclear, é utilizada para transcrever o genoma mitocondrial de 16 kb, que contém duas sequências promotoras relacionadas, uma para cada fita de genoma circular. Cada fita é transcrita em sua totalidade e os transcritos mitocondriais são então processados para gerar os vários RNAms, RNAts e RNArs mitocondriais individuais.

DIVERSIDADE FUNCIONAL CRESCENTE DAS PROTEÍNAS

Muitas proteínas passam por extensos empacotamentos e processamentos pós-traducionais à medida que adotam a sua forma funcional final (Cap. 12). A cadeia polipeptídica, que é o produto de tradução primário, dobra sobre si mesma e forma ligações intramoleculares, criando uma estrutura tridimensional específica, que é determinada pela sequência de aminoácidos em si. Duas ou mais cadeias polipeptídicas, produtos do mesmo gene ou de genes diferentes, podem combinar-se formando um complexo multiproteico único. Por exemplo, duas cadeias de α-globina e duas cadeias de β-globina associam-se de forma não covalente para formar uma molécula de hemoglobina tetramérica (Cap. 11). Os produtos proteicos podem também ser quimicamente modificados, por exemplo, pela adição de grupos metil, fosfatos ou carboidratos em locais específicos. Essas modificações podem ter influência significativa na função ou na abundância da proteína modificada. Outras modificações podem envolver a clivagem da proteína, tanto para remover sequências aminoterminais específicas depois de elas terem funcionado para direcionar uma proteína a sua localização correta dentro da célula (p. ex., proteínas que funcionam dentro da mitocôndria) ou para dividir a molécula em cadeias polipeptídicas menores. Por exemplo, as duas cadeias que compõem a insulina madura, uma com tamanho de 21 e outra de 30 aminoácidos, são originalmente parte de um produto de tradução primário de 82 aminoácidos chamado de proinsulina.

EXPRESSÃO GÊNICA EM AÇÃO

O fluxo de informações descritas nas seções anteriores pode ser mais bem compreendido usando-se como referência um determinado gene bem estudado, o gene da β-globina. A cadeia de β-globina é um polipeptídeo de 146 aminoácidos, codificada por um gene que ocupa aproximadamente 1,6 kb no braço curto do cromossomo 11. O gene possui três éxons e dois íntrons (Fig. 3-4). O gene da β-globina, assim como outros genes do *cluster* de β-globina (Fig. 3-2), é transcrito na direção do centrômero para o telômero. A orientação, no entanto, é distinta para diferentes genes no genoma e depende de qual fita da dupla-hélice cromossômica é a fita codificante para um determinado gene.

As sequências de DNA necessárias para o início preciso da transcrição do gene da β-globina estão localizadas no promotor dentro de cerca de 200 pb a montante do local de início da transcrição. A sequência do DNA de dupla-fita dessa região do gene de β-globina, a sequência de RNA correspondente e a sequência traduzida dos primeiros 10 aminoácidos são representadas na Figura 3-6 para ilustrar as relações entre esses três níveis de informação. Como mencionado anteriormente, é a fita de 3′ a 5′ do DNA que serve como molde e é, na verdade, transcrita, mas é a fita de

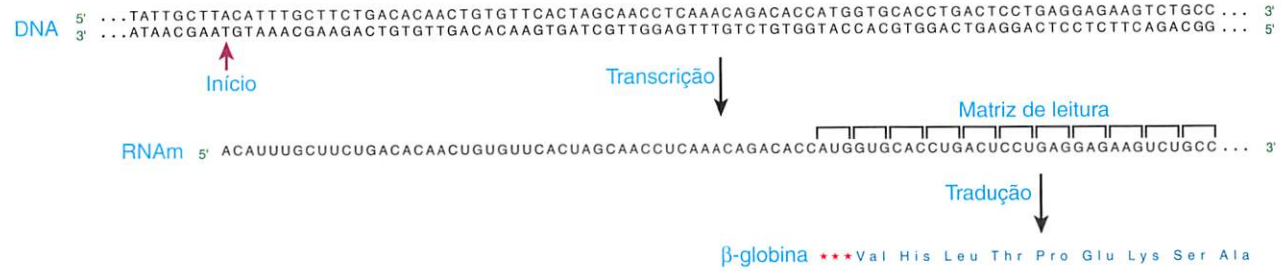

Figura 3-6 **Estrutura e sequência de nucleotídeos da extremidade 5′ do gene de β-globina humana no braço curto do cromossomo 11.** A transcrição da fita de 3′ a 5′ (*inferior*) começa no sítio de início indicado, produzindo o RNA mensageiro de β-globina (RNAm). A matriz de leitura tradicional é determinada pelo códon iniciador AUG (***); códons subsequentes especificando aminoácidos são indicados em *azul*. As outras duas matrizes potenciais não são utilizadas.

5′ a 3′ do DNA que corresponde diretamente à sequência 5′ a 3′ do RNAm (e, de fato, é idêntica a ela, exceto que U é substituído por T). Por causa dessa correspondência, a fita de DNA de 5′ a 3′ de um gene (i.e., a fita que *não* é transcrita) é a fita geralmente relatada na literatura científica ou nos bancos de dados.

De acordo com essa convenção, a sequência completa de aproximadamente 2,0 kb do cromossomo 11 que inclui o gene da β-globina é mostrada na Figura 3-7. (É sensato refletir que uma cópia impressa de todo o genoma humano nessa escala exigiria mais de 300 livros do tamanho deste!) Dentro desses 2,0 kb está contida a maioria dos elementos, mas não todos, de sequência necessários para codificar e regular a expressão desse gene. Muitas das características estruturais importantes do gene da β-globina estão indicadas na Figura 3-7, incluindo elementos de sequências promotoras conservados, os limites íntron-éxon, 5′ e 3′ UTRs, sítios de *splicing* de RNA, os códons iniciador e de término e o sinal de poliadenilação, todos os quais são conhecidos por serem mutados em vários defeitos hereditários do gene da β-globina (Cap. 11).

Início da Transcrição

O promotor da β-globina, como muitos outros promotores de genes, consiste em uma série de elementos funcionais relativamente curtos que interagem com proteínas reguladoras específicas (genericamente chamadas de **fatores de transcrição**) que controlam a transcrição, incluindo, no caso dos genes de globina, aquelas proteínas que restringem a expressão desses genes em células eritroides, as células em que a hemoglobina é produzida. Há bem mais de 1.000 fatores de transcrição de ligação ao DNA sequência-específicos no genoma, sendo que alguns deles são ubíquos em sua expressão, enquanto outros são específicos para o tipo celular ou tecido.

Uma sequência promotora importante encontrada em muitos dos genes, mas não em todos, é a **TATA box**, uma região conservada rica em adeninas e timinas que está, aproximadamente, 25 a 30 pb a montante do sítio de início da transcrição (Figs. 3-4 e 3-7). A TATA box parece ser importante para determinar a posição do início de transcrição, que no gene de β-globina está aproximadamente 50 pb a montante do sítio de início da tradução (Fig. 3-6). Então, nesse gene, existem aproximadamente 50 pb da sequência na extremidade 5′ que são transcritos mas não são traduzidos; em outros genes, a 5′ UTR pode ser muito mais longa e pode ser interrompida por um ou mais íntrons. Uma segunda região conservada, a chamada CAT box (na verdade CCAAT), está a poucas dúzias de pares de bases mais a montante (Fig. 3-7). Tanto mutações experimentalmente induzidas como as de ocorrência natural nesses elementos de sequência, bem como em outras sequências reguladoras ainda mais a montante, levam a uma redução acentuada no nível da transcrição, demonstrando assim a importância desses elementos para a expressão gênica normal. Muitas mutações nesses elementos reguladores têm sido identificadas em pacientes com o distúrbio da hemoglobina β-talassemia (Cap. 11).

Nem todos os promotores de genes contêm os dois elementos específicos que acabamos de descrever. Em particular, os genes que são constitutivamente expressos na maioria ou em todos os tecidos (os chamados genes de manutenção — *housekeeping genes*) muitas vezes não têm os boxes CAT e TATA, que são mais típicos dos genes tecido-específicos. Os promotores de muitos genes de manutenção contêm uma alta proporção de citosinas e guaninas em relação ao DNA circundante (veja o promotor do gene *BRCA1* do câncer de mama na Fig. 3-4). Tais promotores ricos em CG são muitas vezes localizados em regiões do genoma chamadas de **ilhas CpG**, assim denominadas por causa da concentração surpreendentemente alta do dinucleotídeo 5′-CpG-3′ (o *p* representa o grupo fosfato entre bases adjacentes; veja a Fig. 2-3), que se destaca de um panorama genômico mais geral rico em AT. Acredita-se que alguns dos elementos de sequência rica em CG encontrados nesses promotores servem como sítios de ligação para fatores de transcrição específicos. As ilhas de CpG também são importantes porque elas são alvos de **metilação de DNA**. A metilação extensa do DNA nas ilhas CpG está geralmente associada à repressão da transcrição gênica, como discutiremos mais adiante no contexto da cromatina e do seu papel no controle da expressão gênica.

A transcrição pela RNA polimerase II (RNA pol II) é sujeita à regulação em múltiplos níveis, incluindo a ligação

Figura 3-7 **Sequência de nucleotídeos do gene da β-globina humana completo.** É mostrada a sequência da fita de 5′ a 3′ do gene. As áreas *acastanhadas* com letras maiúsculas representam sequências exônicas que correspondem ao RNAm maduro. As letras minúsculas indicam íntrons e sequências flanqueadoras. As sequências CAT e TATA box na região flanqueadora 5′ são indicadas na cor *marrom*. Os dinucleotídeos GT e AG, importantes para o *splicing* de RNA nas junções íntron-éxon, e o sinal AATAAA, importante para a adição de uma cauda poliA, estão também realçados. O códon iniciador ATG (AUG no RNAm) e o códon de parada TAA (UAA no RNAm) são mostrados em letras *vermelhas*. A sequência de aminoácidos de β-globina é mostrada acima da sequência codificante; as abreviações de três letras na Tabela 3-1 são usadas aqui. *Veja Fontes & Agradecimentos.*

com o promotor, o início da transcrição, o desenrolamento da dupla-hélice de DNA para expor a fita-molde e o alongamento à medida que a RNA pol II se move ao longo do DNA. Embora alguns genes silenciados sejam desprovidos de ligação da RNA pol II no conjunto, compatível com a sua incapacidade de serem transcritos em um determinado tipo celular, outros possuem RNA pol II preparada bidirecionalmente no sítio de início da transcrição, talvez como um meio de transcrição afinado em resposta a determinados sinais celulares.

Além das sequências que constituem um promotor em si, existem outros elementos de sequência que podem alterar significativamente a eficiência da transcrição. As sequências mais bem caracterizadas dessas "ativadoras" são chamadas de **acentuadores**. Os acentuadores são elementos de sequência que podem atuar à distância de um

gene (geralmente várias ou mesmo centenas de quilobases de distância) para estimular a transcrição. Ao contrário dos promotores, os acentuadores são independentes tanto em posição como em orientação e podem estar localizados a 5′ ou 3′ do sítio de início da transcrição. Elementos específicos de acentuadores funcionam apenas em determinados tipos celulares e, portanto, parecem estar envolvidos no estabelecimento da especificidade tecidual ou no nível de expressão de muitos genes, em conjunto com um ou mais fatores de transcrição. No caso do gene da β-globina, vários acentuadores tecido-específicos estão presentes tanto dentro do próprio gene como nas suas regiões flanqueadoras. A interação de acentuadores com proteínas reguladoras específicas leva a níveis aumentados de transcrição.

A expressão normal do gene da β-globina durante o desenvolvimento também requer sequências mais distantes, chamadas de **região controladora de *locus*** (**RCL**), localizadas a montante do gene de ε-globina (Fig. 3-2), que são essenciais para o estabelecimento do contexto adequado da cromatina necessário para a expressão de alto nível apropriada. Como esperado, as mutações que interrompem ou eliminam o acentuador ou as sequências de RCL interferem ou impedem a expressão do gene da β-globina (Cap. 11).

Splicing de RNA

O transcrito de RNA primário do gene de β-globina contém dois íntrons, de cerca de 100 e 850 pb de tamanho, que precisam ser removidos, e os segmentos remanescentes de RNA unidos para formar o RNAm maduro. O processo de ***splicing* de RNA**, descrito em linhas gerais anteriormente, é minucioso e altamente eficiente; acredita-se que 95% dos transcritos de β-globina sofram *splicing* com precisão, produzindo RNAm funcional de globina. As reações de *splicing* são guiadas por sequências específicas no transcrito de RNA primário em ambas as extremidades, 5′ e 3′, dos íntrons. A sequência 5′ consiste em nove nucleotídeos, dos quais dois (o dinucleotídeo GT [GU no transcrito de RNA] localizado no íntron imediatamente adjacente ao sítio de *splicing*) praticamente não variam entre sítios de *splicing* de diferentes genes (Fig. 3-7). A sequência 3′ consiste em aproximadamente uma dúzia de nucleotídeos, dos quais, mais uma vez, dois — o AG localizado imediatamente a 5′ do limite íntron-éxon — são obrigatórios para o *splicing* normal. Os locais de *splicing* por si sós não estão relacionados com a matriz de leitura de um determinado RNAm. Em algumas circunstâncias, como no caso do íntron 1 do gene de β-globina, o íntron, na verdade, divide um códon específico (Fig. 3-7).

O significado clínico do *splicing* de RNA é ilustrado pelo fato de que mutações dentro das sequências conservadas nos limites íntron-éxon comumente prejudicam o *splicing* de RNA, com uma redução concomitante da quantidade normal de RNAm de β-globina maduro; mutações nos dinucleotídeos GT ou AG mencionados anteriormente invariavelmente eliminam o *splicing* normal do íntron que contém a mutação. Mutações de sítios de *splicing* representativas, identificadas em pacientes com β-talassemia, são discutidas em detalhes no Capítulo 11.

Splicing Alternativo

Como discutido anteriormente, quando os íntrons são removidos do transcrito de RNA primário pelo *splicing* de RNA, os éxons remanescentes sofrem *splicing* juntos, gerando o RNAm maduro final. No entanto, para a maioria dos genes, o transcrito primário pode seguir múltiplas vias alternativas de *splicing*, o que leva à síntese de múltiplos RNAms relacionados porém diferentes, sendo que cada um dos quais pode ser subsequentemente traduzido para gerar produtos proteicos diferentes (Fig. 3-1). Alguns desses eventos alternativos são altamente tecido- ou tipo celular-específicos e, na medida em que tais eventos são determinados pela sequência primária, eles estão sujeitos à variação alélica entre indivíduos diferentes. Quase todos os genes humanos sofrem *splicing* alternativo em algum grau e estima-se que há uma média de dois ou três transcritos alternativos por gene no genoma humano, expandindo, assim, enormemente o conteúdo de informações do genoma humano para além dos 20.000 genes codificantes de proteínas. A regulação do *splicing* alternativo parece desempenhar um papel particularmente impressionante durante o desenvolvimento neuronal, no qual pode contribuir para a geração de níveis elevados de diversidade funcional necessária no sistema nervoso. Consistente com isso, a suscetibilidade a um número de condições neuropsiquiátricas tem sido associada a mudanças ou ruptura dos padrões de *splicing* alternativo.

Poliadenilação

O RNAm maduro de β-globina contém aproximadamente 130 pb de material de 3′ não traduzido (o 3′ UTR) entre o códon de parada e o local da cauda de poliA (Fig. 3-7). Como em outros genes, a clivagem da extremidade 3′ do RNAm e a adição da cauda poliA são controladas, pelo menos em parte, por uma sequência de AAUAAA de aproximadamente 20 pb antes do sítio de poliadenilação. As mutações nesse sinal de poliadenilação em pacientes com β-talassemia documentam a importância desse sinal para a clivagem adequada de 3′ e a poliadenilação (Cap. 11). A 3′ UTR de alguns genes pode alcançar até vários kb de tamanho. Outros genes possuem vários sítios de poliadenilação alternativos, sendo que a seleção de um deles pode influenciar a estabilidade do RNAm resultante e, assim, o nível do estado de estabilidade de cada RNAm.

Edição de RNA e Diferenças de Sequência de RNA-DNA

Achados recentes sugerem que o princípio conceitual subjacente ao dogma central — de que o RNA e as sequências de proteínas refletem a sequência genômica subjacente — nem sempre é verdadeiro. A edição de RNA para alterar a sequência de nucleotídeos do RNAm foi demonstrada em vários organismos, incluindo os humanos. Esse processo

envolve a desaminação de adenosina em sítios específicos, convertendo um A na sequência de DNA em inosina no RNA resultante; este é então lido pela maquinaria de tradução como um G, levando a alterações na expressão gênica e função proteica, especialmente no sistema nervoso. Diferenças de RNA-DNA mais difundidas envolvendo outras bases (com alterações correspondentes na sequência de aminoácidos codificada) também têm sido relatadas, em níveis que variam entre os indivíduos. Embora o(s) mecanismo(s) e a relevância clínica desses eventos permaneçam controversos, eles ilustram a existência de uma gama de processos capazes de aumentar a diversidade de transcritos e do proteoma.

ASPECTOS EPIGENÉTICOS E EPIGENÔMICOS DA EXPRESSÃO GÊNICA

Dada a variedade de funções e destinos que células diferentes em qualquer organismo devem adotar durante sua vida útil, é evidente que nem todos os genes no genoma podem ser ativamente expressos em todas as células em todos os momentos. Assim como a conclusão do Projeto Genoma Humano foi importante para contribuir para a nossa compreensão da biologia humana e de doenças, identificar as sequências e as características genômicas que orientam os aspectos de desenvolvimento, espaciais e temporais da expressão gênica continua sendo um desafio formidável. Várias décadas de trabalho em biologia molecular definiram elementos reguladores críticos para muitos genes individuais, como vimos na seção anterior, e uma atenção mais recente tem sido direcionada para a realização desses estudos do genoma em uma escala ampla.

No Capítulo 2, apresentamos os aspectos gerais da cromatina que empacotam o genoma e seus genes em todas as células. Aqui, vamos explorar as características específicas da cromatina que estão associadas com genes ativos ou reprimidos como um passo para identificar o código regulador para expressão do genoma humano. Tais estudos concentram-se em alterações reversíveis no ambiente da cromatina como determinantes da função gênica, em vez de alterações na sequência do genoma por si, e são, portanto, chamadas de *epi*genéticas ou, quando consideradas no contexto do genoma como um todo, de *epi*genômicas (do grego *epi*, sobre ou em cima).

O campo da **epigenética** está crescendo rapidamente e consiste no estudo das mudanças hereditárias na função celular ou expressão gênica que podem ser transmitidas de uma célula para outra (e até mesmo de geração a geração), como resultado de sinais moleculares baseados na cromatina (Fig. 3-8). Estados epigenéticos complexos podem ser estabelecidos, mantidos e transmitidos por uma variedade de mecanismos: modificações no DNA, tais como a **metilação do DNA**; inúmeras **modificações de histona** que alteram o empacotamento da cromatina ou o acesso a ela; e substituição de **variantes de histona** especializadas que marcam a cromatina associada a sequências ou regiões particulares no genoma. Essas mudanças de cromatina podem ser altamente dinâmicas e transitórias, capazes de responder rapidamente e de maneira sensível às necessidades de mudança na célula, ou podem ser de longa duração, capazes de serem transmitidas através de múltiplas divisões celulares ou mesmo para gerações subsequentes. Em ambos os casos, o conceito fundamental é que mecanismos epigenéticos *não* alteram a sequência de DNA subjacente e isso os distinguem de mecanismos genéticos, os quais são baseados na sequência. Juntas, as marcas epigenéticas e a sequência de DNA compõem o conjunto de sinais que orientam o genoma a expressar seus genes no momento certo, no lugar certo e nas quantidades certas.

Cada vez mais, evidências apontam que as alterações epigenéticas tenham um papel em doenças humanas em resposta a influências ambientais ou de estilo de vida. A natureza dinâmica e reversível das mudanças epigenéticas possibilita um nível de adaptabilidade ou plasticidade que excede em muito a capacidade da sequência de DNA isoladamente e, portanto, é relevante tanto para as origens como para o tratamento potencial da doença. Vários projetos epigenômicos em larga escala (semelhantes ao Projeto de Genoma Humano original) foram iniciados para catalogar os sítios de metilação do DNA em larga escala no genoma (o chamado metiloma), para avaliar ambientes de CpG ao longo do genoma, para descobrir novas variantes de histonas e padrões de modificação em vários tecidos e para documentar o posicionamento de nucleossomos ao longo do genoma em diferentes tipos celulares e em amostras tanto de indivíduos assintomáticos como daqueles com câncer ou outras doenças. Essas análises são parte de um esforço amplo (o chamado **Projeto ENCODE**, para *Enc*yclopedia *o*f *D*NA *E*lements) para explorar padrões epigenéticos na cromatina em larga escala no genoma, a fim de compreender melhor o controle da expressão gênica em diferentes tecidos ou estados de doença.

Metilação do DNA

A metilação do DNA envolve a modificação de bases de citosina por metilação do carbono na quinta posição no anel de pirimidina (Fig. 3-9). A metilação extensa do DNA é uma marca de genes reprimidos e é um mecanismo difundido e associado ao estabelecimento de programas específicos de expressão gênica durante a diferenciação e o desenvolvimento celular. Tipicamente, a metilação do DNA ocorre no C de dinucleotídeos CpG (Fig. 3-8) e inibe a expressão gênica pelo recrutamento de proteínas específicas de ligação a metil-CpG, que, por sua vez, recrutam enzimas de modificação da cromatina para silenciar a transcrição. A presença de 5-metilcitosina (5-mC) é considerada uma marca epigenética estável que pode ser transmitida fielmente através da divisão celular; no entanto, estados alterados de metilação são frequentemente observados no câncer, com hipometilação de segmentos genômicos grandes ou com hipermetilação regional (particularmente em ilhas de CpG) em outros (Cap. 15).

Uma desmetilação extensa ocorre durante o desenvolvimento das células germinativas e nas fases iniciais de

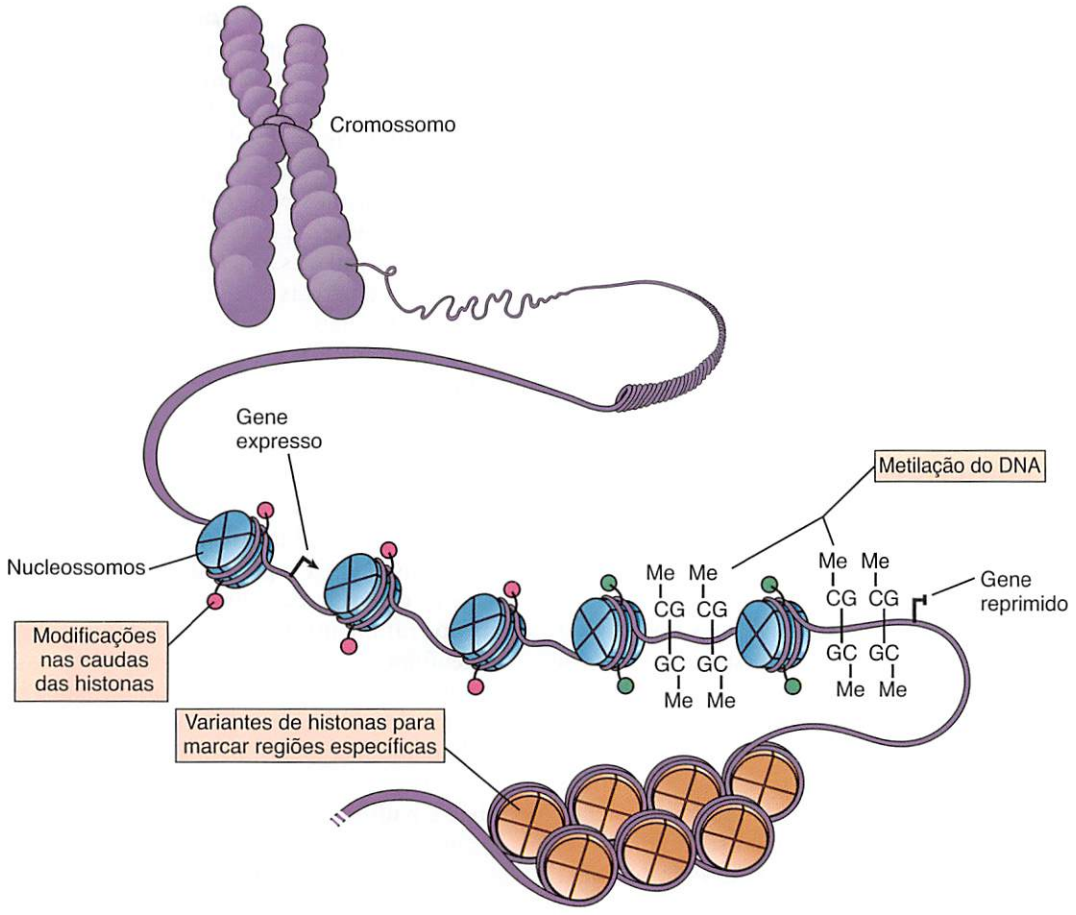

Figura 3-8 Representação esquemática da cromatina e os três principais mecanismos epigenéticos: a metilação de DNA em dinucleotídeos CpG, associada à repressão gênica; várias modificações (indicadas por cores diferentes) nas caudas das histonas, associadas tanto com expressão quanto com repressão gênica; e diversas variantes de histonas que marcam as regiões específicas do genoma, associadas a funções específicas necessárias para a estabilidade cromossômica ou integridade do genoma. Não está em escala.

desenvolvimento embrionário, compatível com a necessidade de "redefinir" o ambiente da cromatina e restaurar a totipotência ou pluripotência do zigoto e de várias populações de células-tronco. Embora os detalhes ainda não sejam totalmente compreendidos, essas etapas de reprogramação parecem envolver a conversão enzimática de 5-mC a 5-hidroximetilcitosina (5-hmC; veja a Fig. 3-9), como um provável intermediário na desmetilação do DNA. Em geral, os níveis de 5-mC são estáveis ao longo dos tecidos adultos (aproximadamente 5% de todas as citocinas), enquanto os níveis de 5-hmC são muito menores e muito mais variáveis (0,1% a 1% de todas as citosinas). Curiosamente, embora a 5-hmC seja bem difundida no genoma, seus níveis mais altos são encontrados em regiões reguladoras conhecidas, sugerindo um possível papel na regulação dos promotores específicos e acentuadores.

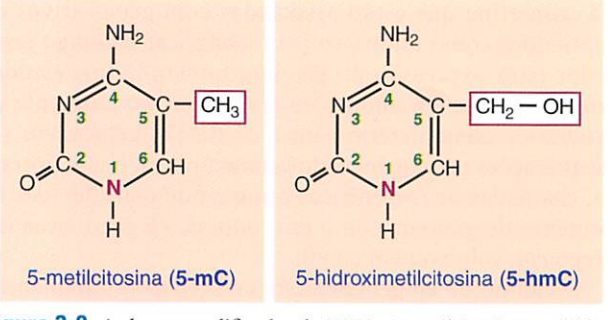

Figura 3-9 **As bases modificadas do DNA, 5-metilcitosina e 5-hidroximetilcitosina.** Compare com a estrutura de citosina na Figura 2-2. Os grupamentos metil e hidroximetil estão marcados em *roxo*. Os átomos nos anéis de pirimidina estão numerados de 1 a 6 para indicar o carbono 5.

Modificações de Histona

Uma segunda classe de sinais epigenéticos consiste em uma lista extensa de modificações em qualquer dos tipos principais de histonas, H2A, H2B, H3 e H4 (Cap. 2). Essas modificações incluem a metilação, a fosforilação, a acetilação das histonas e outros, ocorrendo em resíduos de aminoácidos específicos, localizados principalmente nas "caudas" N-terminais de histonas, que se estendem para fora a partir do centro do nucleossomo (Fig. 3-8). Acredita-se que essas modificações epigenéticas influenciem a expressão gênica,

afetando a compactação da cromatina ou sua acessibilidade e sinalizando complexos de proteínas que — dependendo da natureza do sinal — ativam ou silenciam a expressão gênica naquele local. Existem dúzias de sítios modificados que podem ser experimentalmente consultados em larga escala no genoma, utilizando-se anticorpos que reconhecem sítios especificamente modificados — por exemplo, a histona H3 metilada na lisina na posição 9 (metilação de H3K9, usando a abreviação de uma letra K para lisina; veja a Tabela 3-1) ou a histona H3 acetilada na lisina na posição 27 (acetilação H3K27). A primeira é uma marca repressora associada a regiões silenciadas do genoma, ao passo que a última é uma marca para regiões reguladoras ativas.

Padrões específicos de modificações diferentes de histona estão associados a promotores, a acentuadores ou ao conjunto de genes em diferentes tecidos e tipos celulares. O Projeto ENCODE, apresentado anteriormente, examinou 12 das modificações mais comuns em quase 50 tipos celulares diferentes e integrou os perfis de cromatina individuais a supostos atributos funcionais em mais de metade do genoma humano. Esse achado sugere que uma porção muito maior do genoma desempenha um papel, direta ou indiretamente, na determinação dos padrões variados de expressão gênica que distinguem os tipos celulares do que havia sido previamente inferido, a partir do fato de que menos de 2% do genoma é "codificante" em um sentido tradicional.

Variantes de Histona

As modificações da histona discutidas envolvem modificações das principais histonas em si, que são todas codificadas por *clusters* multigênicos em poucos locais no genoma. Ao contrário, as muitas dezenas de variantes de histona são produtos de genes completamente diferentes, localizados em partes diferentes do genoma, e suas sequências de aminoácidos são distintas das histonas canônicas, apesar de estarem relacionadas.

Diferentes variantes de histonas estão associadas a diferentes funções e substituem — completa ou parcialmente — o membro relacionado das histonas principais encontradas nos nucleossomos típicos para gerar estruturas de cromatina especializadas (Fig. 3-8). Algumas variantes marcam regiões específicas ou *loci* no genoma com funções altamente especializadas; por exemplo, a histona CENP-A é uma variante de histona relacionada com a H3, que é encontrada exclusivamente em centrômeros funcionais no genoma e contribui para as características essenciais da cromatina centromérica que marcam a localização de cinetocoros ao longo da fibra do cromossomo. Outras variantes são mais transitórias e marcam regiões do genoma com atributos particulares; por exemplo, H2A.X é uma histona variante de H2A envolvida na resposta a danos ao DNA para marcar regiões do genoma que requerem reparo do DNA.

Arquitetura da Cromatina

Em contraste com a impressão que se tem ao visualizar o genoma como uma cadeia linear de sequência (Fig. 3-7), o genoma adota uma disposição altamente ordenada e dinâmica dentro do espaço do núcleo, correlacionada com e provavelmente guiada por sinais epigenéticos e epigenômicos que acabamos de discutir. Essa paisagem tridimensional é altamente preditiva do mapa de todas as sequências expressas em qualquer tipo celular determinado (**transcriptoma**) e reflete mudanças dinâmicas na arquitetura da cromatina em diferentes níveis (Fig. 3-10). Em primeiro lugar, grandes domínios cromossômicos (até milhões de pares de bases em tamanho) podem exibir padrões coordenados de expressão gênica em nível cromossômico, envolvendo interações dinâmicas entre diferentes pontos de contato intra e intercromossômicos no interior do núcleo. Em um nível mais aprimorado, avanços técnicos para mapear e sequenciar pontos de contato ao longo do genoma no contexto do espaço tridimensional apontaram para alças ordenadas de cromatina que posicionam e orientam os genes com precisão, expondo ou bloqueando regiões reguladoras críticas para acesso da RNA pol II, de fatores de transcrição e de outros reguladores. Por último, padrões específicos e dinâmicos de posicionamento dos nucleossomos diferem entre os tipos celulares e tecidos em face às mudanças de indícios ambientais e de desenvolvimento (Fig. 3-10). As propriedades biofísicas, epigenômicas e/ou genômicas que facilitam ou especificam o empacotamento ordenado e dinâmico de cada cromossomo durante cada ciclo celular, sem reduzir o genoma a um emaranhado desordenado dentro do núcleo, continuam sendo uma maravilha da engenharia panorâmica.

EXPRESSÃO GÊNICA COMO UMA INTEGRAÇÃO DOS SINAIS GENÔMICOS E EPIGENÔMICOS

A programação de expressão gênica de uma célula inclui um subgrupo específico de aproximadamente 20.000 genes codificantes de proteínas no genoma que são transcritos e traduzidos ativamente em seus respectivos produtos funcionais, o subconjunto dos cerca de 20.000 a 25.000 genes de RNAnc que são transcritos, a quantidade de produtos produzidos e a sequência particular (alelos) daqueles produtos. O perfil de expressão gênica de qualquer célula ou tipo celular em um determinado indivíduo em um determinado momento (quer no contexto do ciclo celular, no desenvolvimento precoce ou durante toda uma vida) e sob um determinado conjunto de circunstâncias (conforme influenciado pelo meio ambiente, estilo de vida ou doença) é, assim, a soma integrada de vários efeitos diferentes, mas inter-relacionados, incluindo os seguintes:

- A sequência primária dos genes, suas variantes alélicas e os seus produtos codificados.
- As sequências reguladoras e o seu posicionamento epigenético na cromatina.
- As interações com os milhares de fatores transcricionais, RNAnc e outras proteínas envolvidas no controle de transcrição, *splicing*, tradução e modificações pós-traducionais.
- A organização do genoma em domínios subcromossômicos.

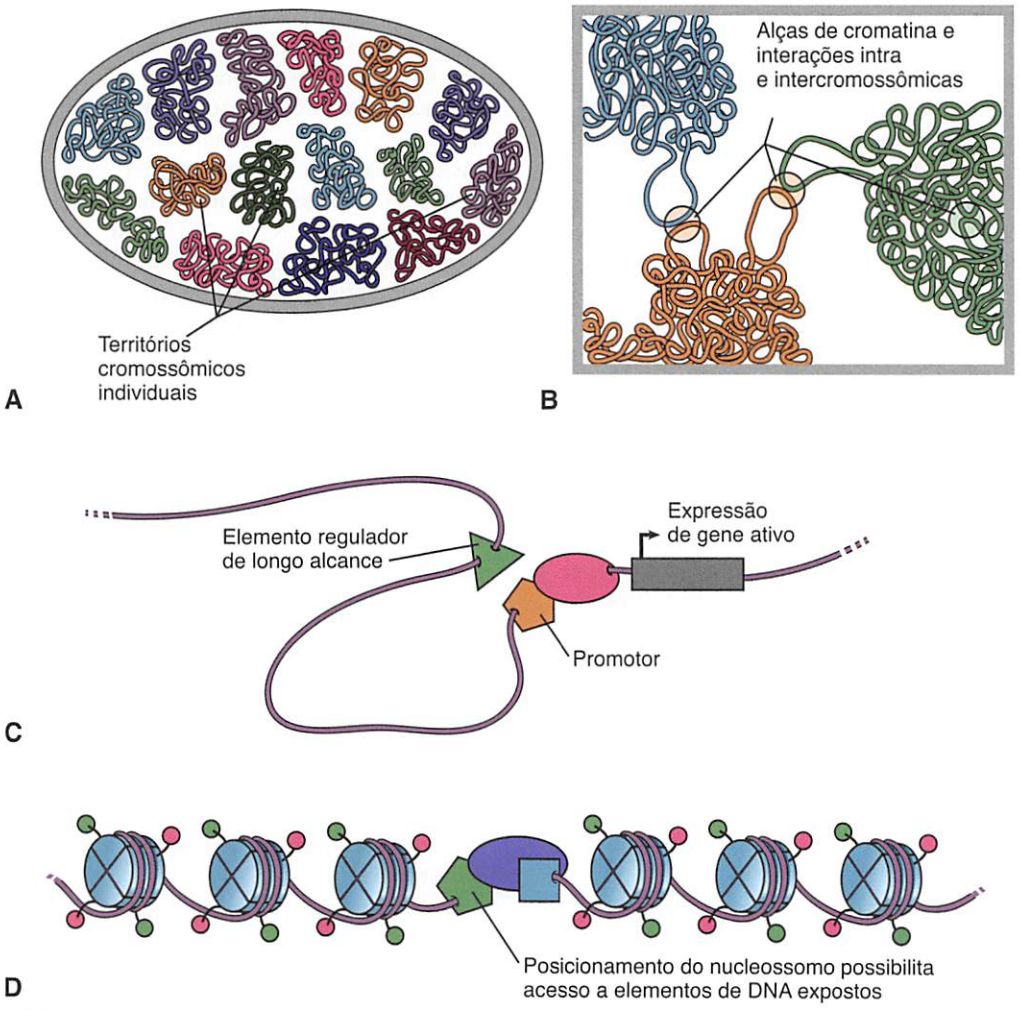

Figura 3-10 **A arquitetura tridimensional e o empacotamento dinâmico do genoma, vistos em níveis crescentes de resolução.** **A,** Dentro do núcleo interfásico, cada cromossomo ocupa um território particular, representado por diferentes cores. **B,** A cromatina é organizada em domínios subcromossômicos grandes dentro de cada território, com alças que trazem determinadas sequências e genes em proximidade uns com os outros, com interações intra e intercromossômicas detectáveis. **C,** As alças trazem elementos reguladores de longo alcance (p. ex., acentuadores ou regiões de controle de *locus*) em associação com promotores, que levam à transcrição ativa e à expressão gênica. **D,** O posicionamento de nucleossomos ao longo da fibra de cromatina promove o acesso a sequências de DNA específicas para a ligação dos fatores de transcrição e outras proteínas reguladoras.

- As interações programadas entre as diferentes partes do genoma.
- O empacotamento tridimensional e dinâmico da cromatina no núcleo.

Todos esses efeitos estão orquestrados de maneira eficiente, hierárquica e altamente programada. Seria de se esperar que a perturbação de qualquer um deles — devido a variação genética, alterações epigenéticas e/ou processos relacionados com doenças — alteraria o programa celular geral e sua produção funcional (Quadro).

DESEQUILÍBRIO ALÉLICO NA EXPRESSÃO GÊNICA

Já se supôs que genes presentes em duas cópias no genoma seriam expressos a partir de ambos os homólogos em níveis comparáveis. No entanto, tornou-se cada vez mais evidente que pode haver um grande desequilíbrio entre os alelos, refletindo tanto a quantidade de variação da sequência no genoma como a interação entre a sequência do genoma e padrões epigenéticos que acabamos de discutir.

PANORAMA EPIGENÉTICO DO GENOMA E MEDICINA

- Os diferentes cromossomos e regiões cromossômicas ocupam territórios característicos dentro do núcleo. A probabilidade de proximidade física influencia a incidência de anormalidades cromossômicas específicas (Caps. 5 e 6).
- O genoma é organizado em domínios de tamanho de megabases com características locais compartilhadas de composição de par de base (i.e., rico em GC ou AT), densidade gênica, momento da replicação na fase S e presença de determinadas modificações de histonas (Cap. 5).
- Os módulos de genes coexpressos correspondem a estágios anatômicos ou de desenvolvimento distintos, por exemplo, no cérebro humano ou na linhagem hematopoiética. Essas redes de coexpressão são reveladas por redes reguladoras compartilhadas e sinais epigenéticos, pelo agrupamento dentro de domínios genômicos e pela sobreposição de padrões de expressão gênica alterada em vários estados de doença.
- Embora os gêmeos monozigóticos compartilhem genomas praticamente idênticos, eles podem ser bastante discordantes para determinados traços, incluindo a suscetibilidade a doenças comuns. Mudanças significativas na metilação do DNA ocorrem durante o tempo de vida desses gêmeos, implicando a regulação epigenética da expressão gênica como uma fonte de diversidade.
- O panorama epigenético pode integrar contribuições genômicas e ambientais à doença. Por exemplo, níveis de metilação de DNA diferenciados correlacionam-se com uma variação subjacente na sequência em *loci* específicos no genoma e, assim, modulam o risco genético para a artrite reumatoide.

No Capítulo 2, introduzimos os achados gerais de que qualquer genoma individual possui dois alelos diferentes em um mínimo de três a cinco milhões de posições ao longo do genoma, distinguindo assim, pela sequência, as cópias herdadas materna e paternalmente daquela posição da sequência (Fig. 2-6). Agora, vamos explorar maneiras pelas quais aquelas diferenças na sequência revelam desequilíbrio alélico na expressão gênica, tanto em *loci* autossômicos como em *loci* do cromossomo X em mulheres.

Pela determinação das sequências de todos os produtos de RNA — o transcriptoma — em uma população de células, pode-se quantificar o nível relativo de transcrição de todos os genes (tanto codificantes como não codificantes de proteínas) que são transcricionalmente ativos nessas células. Considere, por exemplo, o conjunto de genes codificantes de proteínas. Embora uma célula média possa conter aproximadamente 300.000 cópias de RNAm no total, a abundância de RNAms específicos pode diferir em muitas ordens de grandeza; entre genes que estão ativos, a maioria é expressa em níveis baixos (estimados como sendo <10 cópias do RNAm daquele gene por célula), enquanto outros são expressos em níveis muito mais elevados (várias centenas a alguns milhares de cópias daquele RNAm por célula). Apenas em tipos celulares altamente especializados são expressos determinados genes em níveis muito elevados (muitas dezenas de milhares de cópias), correspondendo a uma proporção significativa de todo RNAm nessas células.

Agora, considere um gene expresso com uma variante da sequência que possibilita fazer a distinção entre os produtos de RNA (seja RNAm ou RNAnc) transcritos de cada um de dois alelos, um alelo com um T que é transcrito para produzir um RNA com um A, e o outro alelo com um C que é transcrito para produzir um RNA com um G (Fig. 3-11). Ao sequenciar moléculas de RNA individuais e comparar o número de sequências geradas que contêm um A ou um G naquela posição, pode-se inferir a proporção de transcritos a partir dos dois alelos naquela amostra. Embora a maioria dos genes apresente níveis substancialmente equivalentes de expressão bialélica, análises recentes têm demonstrado uma expressão alélica desigual e generalizada para 5% a 20% dos genes autossômicos no genoma (Tabela 3-2). Para a maioria desses genes, a extensão do desequilíbrio é duplicada ou menor, embora diferenças de até 10 vezes tenham sido observadas para alguns genes. Esse desequilíbrio alélico pode refletir as interações entre a sequência do genoma e a regulação gênica; por exemplo, mudanças na sequência podem alterar a ligação relativa de vários fatores de transcrição ou outros reguladores transcricionais aos dois alelos ou a extensão de metilação do DNA observada nos dois alelos (Tabela 3-2).

Expressão Gênica Monoalélica

Alguns genes, contudo, apresentam uma forma muito mais completa de desequilíbrio alélico, resultando em uma expressão gênica monoalélica (Fig. 3-11). Diversos mecanismos demonstraram ser responsáveis pelo desequilíbrio alélico desse tipo em subgrupos específicos de genes no genoma: rearranjo do DNA, expressão monoalélica aleatória, *imprinting* de origem parental e, para genes no cromossomo X no sexo feminino, inativação do cromossomo X. Suas características distintivas estão resumidas na Tabela 3-2.

Rearranjo Somático

Uma forma de expressão gênica monoalélica altamente especializada é observada nos genes que codificam **imunoglobulinas** e **receptores de células T**, expressos em células B e T, respectivamente, como parte da resposta imunitária. Os anticorpos são codificados na linhagem germinativa por um número relativamente pequeno de genes que, durante o desenvolvimento de células B, são submetidos a um processo único de rearranjo somático. Este processo envolve o corte e a colagem de sequências de DNA nas células precursoras dos linfócitos (mas *não* em quaisquer outras linhagens de células) para reorganizar os genes em células somáticas, gerando uma enorme diversidade de anticorpos. Os rearranjos de DNA altamente orquestrados ocorrem em muitas centenas de quilobases, mas envolvem apenas um dos dois alelos, o qual é escolhido aleatoriamente em qualquer célula B determinada (Tabela 3-2). Assim, a expressão de RNAms maduros para as subunidades da cadeia pesada ou leve de imunoglobulina é exclusivamente monoalélica.

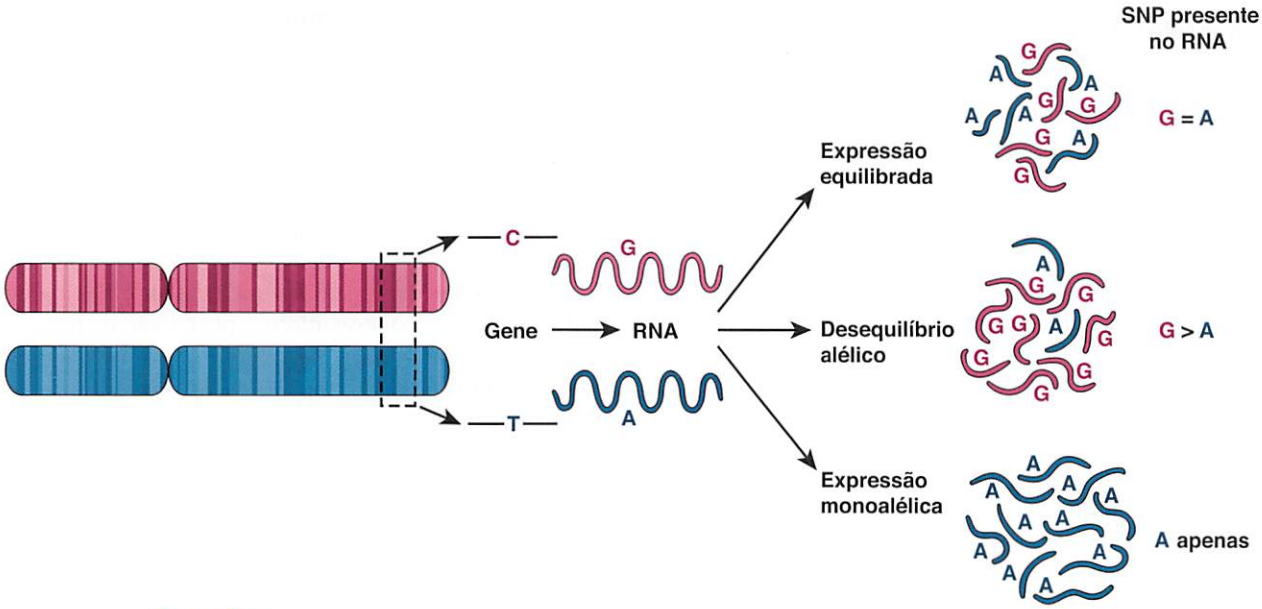

Figura 3-11 **Padrões de expressão alélica para uma sequência gênica com uma variante de DNA transcrita (aqui, um C ou T) para distinguir os alelos.** Como descrito no texto, a abundância relativa de transcritos de RNA dos dois alelos (aqui, carregando um G ou um A) demonstra se o gene apresenta expressão equilibrada (*parte superior*), desequilíbrio alélico (*centro*) ou expressão exclusivamente monoalélica (*parte inferior*). Diferentes mecanismos subjacentes para o desequilíbrio alélico são comparados na Tabela 3-2. SNP, Polimorfismo de nucleotídeo único.

TABELA 3-2 Desequilíbrio Alélico na Expressão Gênica

Tipo	Características	Genes Afetados	Base	Origem no Desenvolvimento
Expressão desequilibrada	Abundância de RNA desigual a partir de dois alelos, devido a variantes de DNA e alterações epigenéticas associadas; geralmente diferença inferior a duas vezes na expressão	5%-20% dos genes autossômicos	Variantes de sequência causam diferentes níveis de expressão nos dois alelos	Embriogênese inicial
Expressão monoalélica				
• Rearranjo somático	Alterações na organização do DNA para produzir um gene funcional em um alelo, mas não em outro	Genes de imunoglobulina, genes de receptores de células T	Escolha aleatória de um alelo	Linhagens de células T e B
• Silenciamento ou ativação alélica aleatória	Expressão a partir de um único alelo em um *locus*, devido ao empacotamento epigenético diferencial no *locus*	Genes de receptores olfativos em neurônios sensoriais; outros genes quimiossensoriais ou do sistema imune; até 10% de todos os genes em outros tipos celulares	Escolha aleatória de um alelo	Tipos celulares específicos
• *Imprinting* genômico	Silenciamento epigenético de alelo(s) na região "imprintada"	> 100 genes com funções no desenvolvimento	Região "imprintada" marcada epigeneticamente de acordo com a origem do progenitor	Linhagem germinativa parental
• Inativação do cromossomo X	Silenciamento epigenético de alelos em um cromossomo X em mulheres	Maioria dos genes ligados ao X em mulheres	Escolha aleatória de um cromossomo X	Embriogênese inicial

Esse mecanismo de rearranjo somático e da expressão gênica monoalélica aleatória também é observado nos genes de receptores de células T na linhagem de células T. Contudo, tal comportamento é exclusivo para essas famílias gênicas e linhagens celulares; o restante do genoma permanece altamente estável ao longo do desenvolvimento e da diferenciação.

Expressão Monoalélica Aleatória

Em contraste com essa forma altamente especializada de rearranjo de DNA, a expressão monoalélica resulta tipicamente da regulação epigenética diferencial dos dois alelos. Um exemplo bem estudado de expressão monoalélica aleatória envolve a família gênica de RO descrita anteriormente (Fig. 3-2). Nesse caso, apenas um único alelo de um

gene de RO é expresso em cada neurônio sensorial olfatório; as muitas centenas de outras cópias da família de RO permanecem reprimidas nessa célula. Outros genes com funções quimiossensoriais ou do sistema imune também apresentam expressão monoalélica aleatória, o que sugere que este mecanismo pode ser geral para aumentar a diversidade de respostas para células que interagem com o mundo exterior. Contudo, esse mecanismo aparentemente não é restrito aos sistemas imunes e sensoriais, porque um subgrupo substancial de todos os genes humanos (5% a 10% em diferentes tipos celulares) tem demonstrado passar por silenciamento alélico aleatório; estes genes estão amplamente distribuídos em todos os autossomos, têm uma ampla gama de funções, e variam nos tipos celulares e tecidos, nos quais a expressão monoalélica é observada.

Imprinting de Origem Parental

Para os exemplos anteriormente descritos, a escolha de qual alelo é expresso não é dependente da origem parental; a cópia materna ou a paterna pode ser expressa em diferentes células e em seus descendentes clonais. Isso distingue formas *aleatórias* de expressão monoalélica de ***imprinting* genômico**, no qual a escolha do alelo a ser expresso é *não aleatória,* sendo determinada unicamente pela origem parental. O *imprinting* é um processo normal que envolve a introdução de marcas epigenéticas (Fig. 3-8) na linhagem germinativa de um dos progenitores, mas não no outro, em locais específicos no genoma. Isto leva à expressão monoalélica de um gene ou, em alguns casos, de múltiplos genes dentro da região "imprintada".

O *imprinting* ocorre durante a gametogênese, antes da fertilização, e marca determinados genes como sendo de origem materna ou paterna (Fig. 3-12). Após a concepção, o *imprinting* de origem parental é mantido em alguns ou todos os tecidos somáticos do embrião, silenciando a expressão gênica no(s) alelo(s) dentro da região "imprintada"; enquanto alguns genes "imprintados" apresentam expressão monoalélica em todo o embrião, outros apresentam *imprinting* tecido-específico, em especial na placenta, com expressão bialélica em outros tecidos. O estado "imprintado" persiste no pós-natal até a idade adulta, através de centenas de divisões celulares, de modo que apenas a cópia materna ou paterna do gene é expressa. No entanto, o *imprinting* deve ser reversível: um alelo de origem paterna, quando herdado por uma mulher, deve ser convertido em sua linhagem germinativa de modo que ela possa passar um *imprint* materno para sua prole. Da mesma maneira, um alelo de origem materna com *imprinting*, quando é herdado por um homem, deve ser convertido em sua linhagem germinativa de maneira que ele possa passá-lo adiante como um alelo de *imprinting* paterno para sua prole (Fig. 3-12). O controle sobre esse processo de conversão parece ser governado por elementos de DNA específicos, chamados de **regiões de controle de *imprinting*** ou **centros de *imprinting*** que estão localizados dentro de regiões "imprintadas" em todo o genoma; embora o seu mecanismo exato de ação não seja conhecido, muitos parecem envolver RNAncs que iniciam a mudança epigenética na cromatina, que, em seguida, se espalha ao longo do cromossomo através da região "imprintada". Notavelmente, embora a região "imprintada" possa abranger mais do que um único gene, essa forma de expressão monoalélica é restrita a um segmento genômico delimitado, tipicamente de algumas centenas de pares de quilobases a algumas megabases de tamanho; isto distingue o *imprinting* genômico tanto da forma mais geral de expressão monoalélica aleatória descrita anteriormente (que parece envolver genes individuais sob controle *locus*-específico), como da inativação do cromossomo X, descrita na próxima seção (que envolve genes ao longo de todo o cromossomo).

Até o momento, cerca de 100 genes "imprintados" foram identificados em muitos autossomos diferentes. O envolvimento desses genes em vários distúrbios cromossômicos é descrito com mais detalhes no Capítulo 6. Para as condições clínicas decorrentes de um único gene "imprintado", tais como a **síndrome de Prader-Willi** (Caso 38) e **síndrome de Beckwith-Wiedemann** (Caso 6), o efeito do *imprinting* genômico nos padrões de herança em heredogramas é discutido no Capítulo 7.

Inativação do Cromossomo X

A base cromossômica para a determinação sexual, introduzida no Capítulo 2 e discutida em mais detalhes no Capítulo 6, resulta em uma diferença de dosagem entre homens e mulheres com relação a genes no cromossomo X. Aqui vamos discutir os mecanismos cromossômicos e moleculares de inativação do cromossomo X, o exemplo mais extenso de expressão monoalélica aleatória no genoma e um mecanismo de **compensação de dose** que resulta no silenciamento epigenético da maioria dos genes em um dos dois cromossomos X nas mulheres.

Em células de mulheres normais, a escolha de qual cromossomo X deve ser inativado é aleatória e mantida em cada linhagem clonal. Assim, as mulheres são mosaico em relação à expressão gênica ligada ao X; algumas células expressam alelos no X herdado de origem paterna, mas não no X de herança materna, ao passo que outras células fazem o oposto (Fig. 3-13). Esse padrão de mosaico da expressão gênica distingue a maioria dos genes ligados ao X dos genes "imprintados", cuja expressão, como acabamos de observar, é determinada estritamente pela origem parental.

Embora o cromossomo X inativo tenha sido primeiramente identificado citologicamente pela presença de uma massa heterocromática (chamada de **corpúsculo de Barr**) em células interfásicas, muitas características epigenéticas distinguem os cromossomos X ativos dos inativos, incluindo a metilação do DNA, modificações de histonas e uma variante de histona específica, a macroH2A, que está particularmente enriquecida na cromatina do X inativo. Além de fornecer conhecimento sobre os mecanismos de inativação de X, essas características podem ser úteis no diagnóstico para identificar cromossomos X inativos em material clínico, como veremos no Capítulo 6.

Embora a inativação do X seja claramente um fenômeno cromossômico, nem todos os genes no cromossomo X apresentam expressão monoalélica em células femininas. A análise extensa da expressão de quase todos

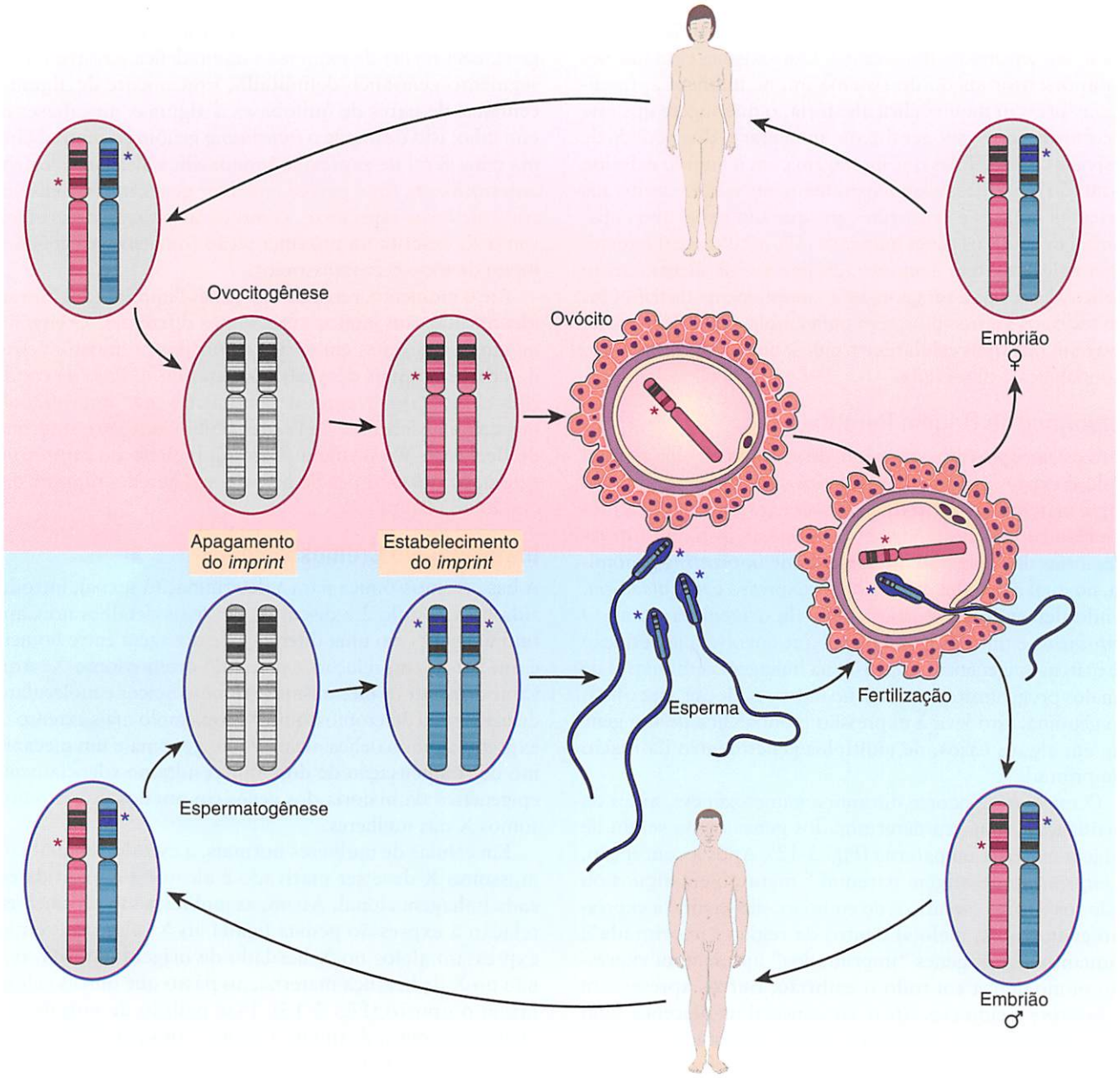

Figura 3-12 ***Imprinting*** **genômico e conversão dos** ***imprints*** **materno e paterno através da gametogênese masculina ou feminina.** Dentro de uma região "imprintada" hipotética em um par de autossomos homólogos, os genes "imprintados" paternalmente estão indicados em *azul*, enquanto um gene "imprintado" maternalmente é indicado em *vermelho*. Após a fecundação, tanto o embrião do sexo masculino como o do sexo feminino têm uma cópia do cromossomo carregando um *imprint* paterno e uma cópia carregando um *imprint* materno. Durante a ovocitogênese (*parte superior*) e espermatogênese (*parte inferior*), os *imprints* são apagados pela remoção das marcas epigenéticas e são estabelecidos novos *imprints* determinados pelo sexo do progenitor na região "imprintada". Os gametas, portanto, realizam *imprint* monoalélico apropriado à origem do progenitor, enquanto as células somáticas em ambos os sexos carregam um cromossomo de cada tipo de *imprint*.

os genes ligados ao X demonstrou que pelo menos 15% dos genes apresentam expressão bialélica e são expressos a partir de cromossomos X ativos e inativos, pelo menos até certo ponto; uma proporção desses apresenta níveis significativamente mais elevados de produção de RNAm em células femininas em relação às células masculinas, sendo candidatos interessantes para explicar traços sexuais dismórficos.

Um subgrupo especial de genes está localizado nos segmentos pseudoautossômicos, que são essencialmente idênticos nos cromossomos X e Y e passam por recombinação durante a espermatogênese (Cap. 2). Esses genes têm duas

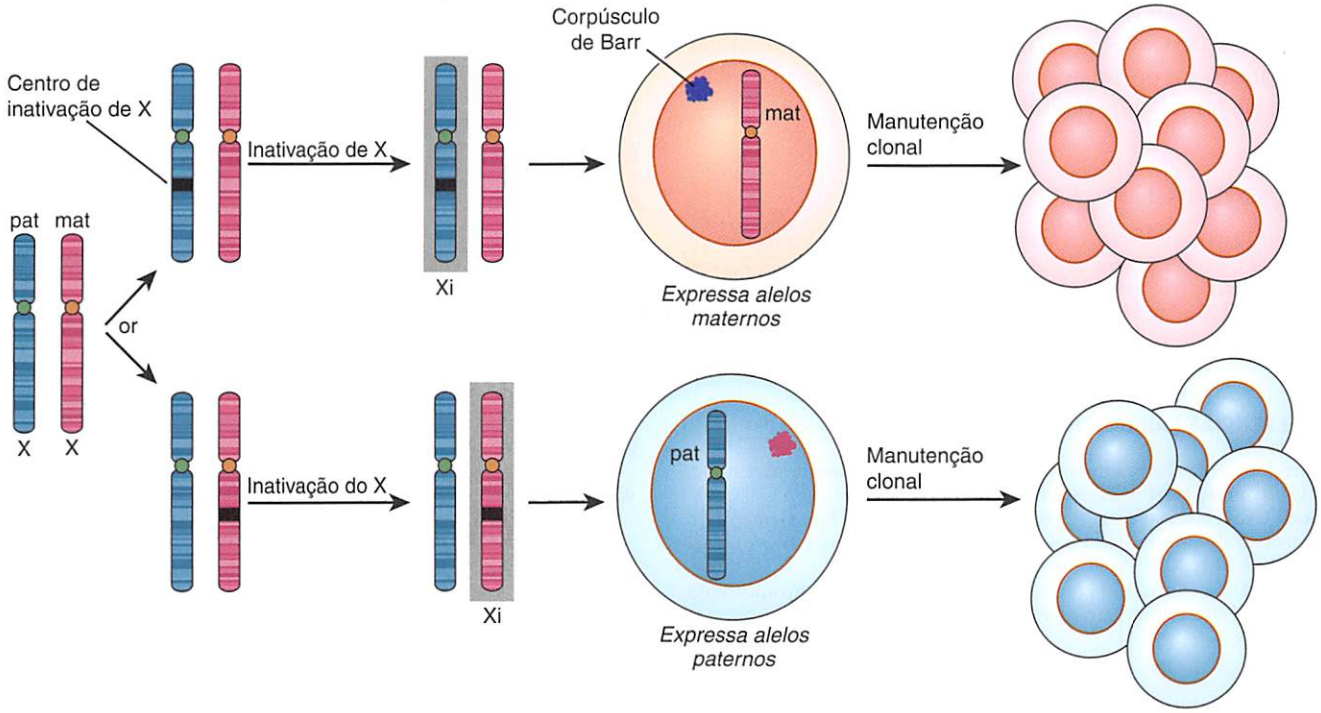

Figura 3-13 **Inativação aleatória do cromossomo X no início do desenvolvimento feminino.** Um pouco depois da concepção de um embrião feminino, tanto os cromossomos X de herança paterna como os de herança paterna (pat e mat, respectivamente) estão ativos. Na primeira semana da embriogênese, um ou outro X é escolhido ao acaso para se tornar o futuro X inativo, por meio de uma série de eventos envolvendo o centro de inativação do X (*quadrado preto*). Esse X torna-se então o X inativo (Xi, indicado pelo *sombreamento*) naquela célula e em sua progênie, e forma o corpúsculo de Barr em núcleos interfásicos. O embrião feminino resultante é, assim, um mosaico clonal de dois tipos de células epigeneticamente determinadas: uma expressa alelos do X materno (células *em rosa*), enquanto a outra expressa alelos do X paterno (células *em azul*). A proporção dos dois tipos de células é determinada aleatoriamente, mas varia entre mulheres normais e entre as mulheres que são portadoras de alelos de doenças ligadas ao X (Caps. 6 e 7).

cópias tanto nas mulheres (duas cópias ligadas ao X) quanto nos homens (uma cópia ligada ao X e uma ligada ao Y) e, portanto, não sofrem inativação do X; como esperado, esses genes apresentam expressão bialélica equilibrada, como se vê na maioria dos genes autossômicos.

Centro de Inativação do X e o Gene XIST. A inativação do X ocorre muito cedo no desenvolvimento embrionário feminino, e a determinação de qual X será designado como X inativo em qualquer célula no embrião é uma escolha aleatória sob controle de um *locus* complexo chamado de **centro de inativação do X**. Essa região contém um gene de RNAnc incomum, o **XIST**, que parece ser um *locus*-mestre regulador importante para a inativação do X. O *XIST* (acrônimo para a expressão em inglês *inactive X* [Xi]-*specific transcripts*) tem a nova característica que é expressa apenas a partir do alelo no X inativo; é transcricionalmente silencioso no X ativo tanto em células masculinas como femininas. Embora o modo exato de ação de *XIST* seja desconhecido, a inativação de X não pode ocorrer na sua ausência. O produto de *XIST* é um RNAnc longo que permanece no núcleo em estreita associação com o cromossomo X inativo.

Outros aspectos e consequências da inativação do cromossomo X serão discutidos no Capítulo 6, no contexto de indivíduos com cromossomos X estruturalmente anormais ou com um número anormal de cromossomos X, e no Capítulo 7, no caso de mulheres que são portadoras de alelos mutantes deletérios para doenças ligadas ao X.

VARIAÇÃO NA EXPRESSÃO GÊNICA E SUA RELEVÂNCIA PARA A MEDICINA

A expressão regulada de genes no genoma humano envolve um conjunto de inter-relações complexas entre diferentes níveis de controle, incluindo a dosagem gênica adequada (controlada por mecanismos de replicação e segregação cromossômica), estrutura gênica, empacotamento de cromatina e regulação epigenética, transcrição, *splicing* de RNA e, para os *loci* codificantes de proteína, estabilidade do RNAm, tradução, processamento e degradação de proteínas. Para alguns genes, oscilações nos níveis do produto do gene funcional, devido à variação hereditária na estrutura de um gene particular ou às alterações induzidas por fatores não genéticos, como a dieta ou o ambiente, são relativamente de pouca importância. Para outros genes, mesmo alterações relativamente menores nos níveis de expressão podem ter consequências clínicas desastrosas, refletindo a importância desses produtos gênicos em vias biológicas específicas. A natureza da variação hereditária na estrutura e na função

dos cromossomos, dos genes e do genoma, combinada à influência dessa variação na expressão de características específicas, é a própria essência da genética médica e molecular e é tratada nos capítulos subsequentes.

REFERÊNCIAS GERAIS

Brown TA: *Genomes*, ed 3, New York, 2007, Garland Science.
Lodish H, Berk A, Kaiser CA, et al: *Molecular cell biology*, ed 7, New York, 2012, WH Freeman.
Strachan T, Read A: *Human molecular genetics*, ed 4, New York, 2010, Garland Science.

REFERÊNCIAS PARA TÓPICOS ESPECÍFICOS

Bartolomei MS, Ferguson-Smith AC: Mammalian genomic imprinting, *Cold Spring Harbor Perspect Biol* 3:1002592, 2011.
Beck CR, Garcia-Perez JL, Badge RM, et al: LINE-1 elements in structural variation and disease, *Annu Rev Genomics Hum Genet* 12:187-215, 2011.
Berg P: Dissections and reconstructions of genes and chromosomes (Nobel Prize lecture), *Science* 213:296-303, 1981.
Chess A: Mechanisms and consequences of widespread random monoallelic expression, *Nat Rev Genet* 13:421-428, 2012.
Dekker J: Gene regulation in the third dimension, *Science* 319:1793-1794, 2008.
Djebali S, Davis CA, Merkel A, et al: Landscape of transcription in human cells, *Nature* 489:101-108, 2012.
ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome, *Nature* 489:57-74, 2012.
Gerstein MB, Bruce C, Rozowsky JS, et al: What is a gene, post-ENCODE? *Genome Res* 17:669-681, 2007.
Guil S, Esteller M: Cis-acting noncoding RNAs: friends and foes, *Nat Struct Mol Biol* 19:1068-1074, 2012.
Heyn H, Esteller M: DNA methylation profiling in the clinic: applications and challenges, *Nature Rev Genet* 13:679-692, 2012.
Hubner MR, Spector DL: Chromatin dynamics, *Annu Rev Biophys* 39:471-489, 2010.
Li M, Wang IX, Li Y, et al: Widespread RNA and DNA sequence differences in the human transcriptome, *Science* 333:53-58, 2011.
Nagano T, Fraser P: No-nonsense functions for long noncoding RNAs, *Cell* 145:178-181, 2011.
Willard HF: The human genome: a window on human genetics, biology and medicine. In Ginsburg GS, Willard HF, editors: *Genomic and personalized medicine*, ed 2, New York, 2013, Elsevier.
Zhou VW, Goren A, Bernstein BE: Charting histone modifications and the functional organization of mammalian genomes, *Nat Rev Genet* 12:7-18, 2012.

PROBLEMAS

1. A sequência de aminoácidos a seguir representa parte de uma proteína. A sequência normal e quatro formas mutantes são mostradas. Consultando a Tabela 3-1, determine a sequência da dupla-fita da seção correspondente do gene normal. Que fita é aquela que a polimerase de RNA "lê"? Qual seria a sequência do RNAm resultante? Que tipo de mutação cada proteína mutante provavelmente representa?
   Normal -lys-arg-his-his-tyr-leu
   Mutante 1 -lys-arg-his-his-cys-leu
   Mutante 2 -lys-arg-ile-ile-ile-
   Mutante 3 -lys-glu-thr-ser-leu-ser-
   Mutante 4 -asn-tyr-leu-

2. Os seguintes itens estão relacionados uns com os outros de maneira hierárquica: cromossomo, par de base, nucleossomo, par de quilobase, íntron, gene, éxon, cromatina, códon, nucleotídeo, promotor. Quais são essas relações?

3. Descreva como se pode esperar que uma mutação em cada uma das seguintes regiões altere ou interfira na função gênica normal, causando doenças humanas: promotor, códon iniciador, sítios de *splicing* nas junções íntron-éxon, uma deleção de um par de base na sequência codificante, códon de parada.

4. A maior parte do genoma humano consiste em sequências que não são transcritas e não codificam produtos gênicos diretamente. Considere maneiras pelas quais os seguintes elementos do genoma podem contribuir para doenças humanas: íntrons, sequências repetitivas *Alu* ou LINE, regiões de controle de *locus*, pseudogenes.

5. Contraste os mecanismos e as consequências do *splicing* de RNA e do rearranjo somático.

6. Considere diferentes maneiras em que mutações ou variações a seguir podem levar a doenças humanas: modificações epigenéticas, metilação do DNA, genes de miRNA, genes de RNAInc.

7. Compare os mecanismos e as consequências do *imprinting* genômico e da inativação do cromossomo X.