Organização do Genoma Humano

9

CONCEITOS PRINCIPAIS

- O genoma humano é subdividido em um grande genoma nuclear, com mais de 26 mil genes, e um genoma mitocondrial circular muito pequeno, com apenas 37 genes. O genoma nuclear é distribuído em 24 moléculas lineares de DNA, uma para cada um dos 24 tipos diferentes de cromossomos humanos.
- Genes humanos geralmente não são estruturas discretas: seus transcritos frequentemente se sobrepõem àqueles de outros genes, às vezes em ambas as fitas.
- A duplicação de genes únicos, de regiões subcromossômicas ou de genomas inteiros deu origem a famílias de genes relacionados.
- Genes são tradicionalmente definidos como sequências que codificam RNA para a consequente síntese de proteínas, mas milhares de genes produzem moléculas de RNA não codificante que podem estar envolvidas em diversas funções.
- RNAs não codificantes muitas vezes regulam a expressão de genes-alvo específicos pelo pareamento de bases com seus transcritos de RNA.
- Algumas cópias de um gene funcional podem sofrer mutações que impedem sua expressão. Estes pseudogenes se originam pela cópia de DNA genômico ou pela cópia de um transcrito de RNA processado em uma sequência de cDNA, que é reintegrada ao genoma (retrotransposição).
- Ocasionalmente, cópias gênicas originadas por retrotransposição mantêm sua função devido à pressão de seleção. Estas são conhecidas como retrogenes.
- Transposons são sequências que se movem de um local para outro no genoma por meio de um mecanismo de corte e colagem ou de cópia e colagem. Retrotransposons fazem uma cópia de cDNA a partir de um transcrito de RNA que então é integrada em um novo local do genoma.
- Grandes arranjos de repetições em *tandem* de alto número de cópias, conhecidos como DNA satélite, estão associados com regiões de heterocromatina altamente condensada e transcricionalmente inativa em cromossomos humanos.

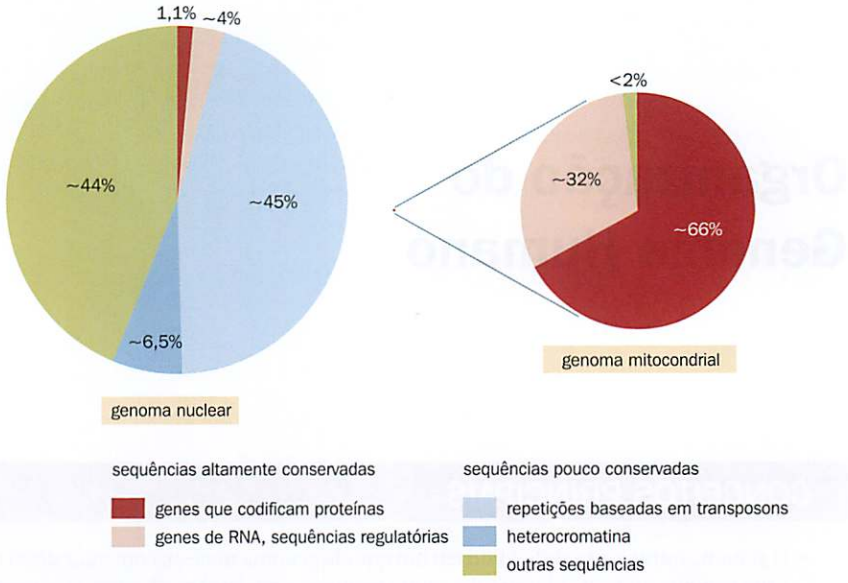


Figura 9.1 Conservação e classes de sequências no genoma nuclear humano e no genoma mitocondrial. Para se obter uma ideia da vasta diferença em escala entre os genomas nuclear (esquerda) e mitocondrial (direita), os pequenos pontos vermelhos no centro representam o equivalente a 25 genomas de DNA mitocondrial (mtDNA) na mesma escala do único genoma nuclear à esquerda. Observe também a profunda diferença entre os dois genomas nas frações de DNA altamente conservado e ainda na fração de DNA não codificante, altamente repetitivo.

O genoma humano consiste em duas partes: um *genoma nuclear* complexo, com mais de 26 mil genes, e um *genoma mitocondrial* bastante simples, com apenas 37 genes (**Figura 9.1**). O genoma nuclear fornece a maior parte da informação genética essencial e é dividido em 23 ou 24 tipos diferentes de moléculas de DNA cromossômico (22 autossomos mais um cromossomo X em mulheres, e um cromossomo Y adicional em homens).

As mitocôndrias possuem seu próprio genoma – um tipo único de DNA circular pequeno –, que codifica alguns dos componentes necessários para a síntese de proteínas nos ribossomos mitocondriais. No entanto, a maioria das proteínas mitocondriais é codificada por genes nucleares e é sintetizada nos ribossomos citoplasmáticos antes de ser importada para a mitocôndria.

Conforme detalhado no Capítulo 10, comparações de sequências com genomas de outros mamíferos e vertebrados indicam que cerca de 5% do genoma humano foi fortemente conservado ao longo da evolução, presumivelmente apresentando importância funcional. Sequências de DNA que codificam proteínas compreendem apenas 1,1% do genoma. Os aproximadamente 4% restantes de sequências conservadas do genoma consistem em sequências de DNA não codificante, incluindo genes cujos produtos finais são moléculas de RNA funcionalmente importantes, e uma variedade de sequências em *cis* que regulam a expressão gênica nos níveis de DNA e RNA. Embora sequências que originam RNA não codificante de proteínas geralmente não apresentem alto grau de conservação ao longo da evolução, algumas sequências regulatórias apresentam um grau de conservação muito maior do que sequências codificantes de proteínas.

Sequências que codificam proteínas frequentemente pertencem a famílias de sequências relacionadas que podem estar organizadas em grupos, em um ou mais cromossomos, ou dispersas ao longo do genoma. Estas famílias se originaram por duplicação gênica durante a evolução. Os mecanismos que originam genes duplicados também dão origem a sequências não funcionais relacionadas a genes (*pseudogenes*).

Uma das grandes surpresas dos últimos anos foi a descoberta de que a transcrição do genoma humano origina dezenas de milhares de transcritos de *RNA não codificante*, incluindo categorias completamente novas de pequenos RNAs regulatórios não identificadas anteriormente na primeira versão do genoma humano, publicada em 2001. Embora se esteja perto de obter uma lista definitiva dos genes humanos que codificam proteínas, o conhecimento sobre os genes de RNA permanece pouco desenvolvido. No entanto, está muito claro que o RNA é funcionalmente muito mais versátil do que suspeitávamos anteriormente. Além de uma lista rapidamente crescente de genes de RNA humanos, também toma-se conhecimento da existência de um enorme número de cópias de pseudogenes de RNA.

Uma grande fração do genoma humano, assim como de outros genomas complexos, é constituída por sequências de DNA não codificante, altamente repetitivas. Uma parte

considerável destas sequências está organizada em repetições em *tandem**, mas a maioria consiste em repetições dispersas que foram copiadas de transcritos de RNA, na célula, pela transcriptase reversa. Há uma crescente conscientização acerca da importância funcional de tais repetições.

Neste capítulo será tratado predominantemente da *arquitetura* do genoma humano. Haverá uma descrição das diferentes classes de sequências de DNA, descrevendo brevemente suas funções, e será abordado como estas sequências estão organizadas no genoma humano. Em capítulos posteriores, serão descritos outros aspectos do genoma humano: como ele se compara a outros genomas e como ele foi moldado pela evolução (Capítulo 10), a variação de sequências de DNA e os polimorfismos (Capítulo 13) e aspectos da expressão dos genes humanos (Capítulo 11).

9.1 ORGANIZAÇÃO GERAL DO GENOMA HUMANO

A sequência de DNA do genoma mitocondrial humano foi publicada em 1981, e uma compreensão detalhada de como o DNA mitocondrial (mtDNA) funciona foi construída a partir de então. O genoma nuclear, mais complexo, foi um desafio muito maior. Um sequenciamento geral do genoma nuclear começou no final dos anos 1990, e em 2004 praticamente toda a porção eucromática do genoma havia sido sequenciada. No entanto, o conhecimento acerca do genoma nuclear ainda é fragmentado. Como será visto a seguir, ainda não se sabe quantos genes há no genoma nuclear, e dados obtidos recentemente estão mudando radicalmente nossa perspectiva sobre como ele está organizado e como se expressa.

O genoma mitocondrial é densamente organizado com informação genética

O genoma mitocondrial humano consiste em um tipo único de molécula de DNA dupla-fita circular, com um tamanho de 16,6 kb. A composição geral de bases é de 44% (G + C), mas as duas fitas do mtDNA são significativamente diferentes em sua composição de bases: a fita pesada (H, *heavy*) é rica em guaninas, enquanto a fita leve (L, *light*) é rica em citosinas. As células contêm, geralmente, milhares de cópias da molécula de mtDNA dupla-fita, mas este número pode variar consideravelmente em diferentes tipos celulares.

Durante a formação do zigoto, um espermatozoide contribui com seu genoma nuclear, mas não com seu genoma mitocondrial, para a formação da célula ovo. Consequentemente, o genoma mitocondrial do zigoto é muitas vezes determinado exclusivamente por aquele originalmente encontrado no óvulo não fecundado. O genoma mitocondrial é, portanto, de herança materna: tanto machos como fêmeas herdam suas mitocôndrias de suas mães, mas os homens não transmitem suas mitocôndrias para as gerações seguintes. Durante a divisão celular mitótica, as várias moléculas de mtDNA de uma célula em divisão segregam de maneira puramente randômica para as duas células-filhas.

Replicação do DNA mitocondrial

A replicação de ambas as fitas H e L é unidirecional e tem início em origens específicas. Embora o DNA mitocondrial seja principalmente dupla-fita, a síntese repetida de um pequeno fragmento da fita H produz uma terceira fita curta, chamada DNA 7S. A fita de DNA 7S pode parear com as bases da fita L e deslocar a fita H, resultando em uma estrutura de fita tripla (**Figura 9.2**). Esta região contém muitas das sequências controladoras do mtDNA (incluindo as principais regiões promotoras) e, portanto, é chamada de *região CR/alça-D* (em que CR significa região de controle, e alça-D significa alça de deslocamento).

A origem de replicação para a fita H está localizada na região CR/alça-D, e aquela para a fita L fica entre dois genes de tRNA (**Figura 9.3**). Apenas após cerca de dois terços da filha H ter sido sintetizada (utilizando a fita L como molde e deslocando a fita H original) é que a origem de replicação para a fita L se torna disponível. Depois disso, a replicação da fita L prossegue na direção oposta, utilizando a fita H como molde.

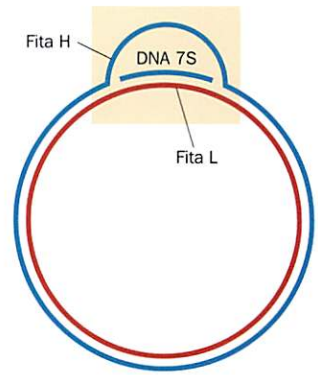


Figura 9.2 Formação de alça-D no DNA mitocondrial. O genoma mitocondrial não é uma simples molécula de dupla-fita de DNA circular. A síntese repetida de um pequeno segmento da fita H (pesada) resulta em uma terceira fita curta (DNA 7S), a qual pode parear com a fita L (leve) e deslocar a fita H, formando uma estrutura de fita tripla local que contém diversas sequências regulatórias importantes e é conhecida como a *região CR/alça-D* (representada por sombreamento e de forma aumentada para maior clareza).

* N. de T.: O termo *tandem*, ou em *tandem*, corresponde ao arranjo de sequências de maneira adjacente, uma após a outra, na mesma direção. O termo foi mantido no idioma original devido à ausência de uma tradução adequada e por se tratar de uma expressão consagrada no jargão da disciplina.

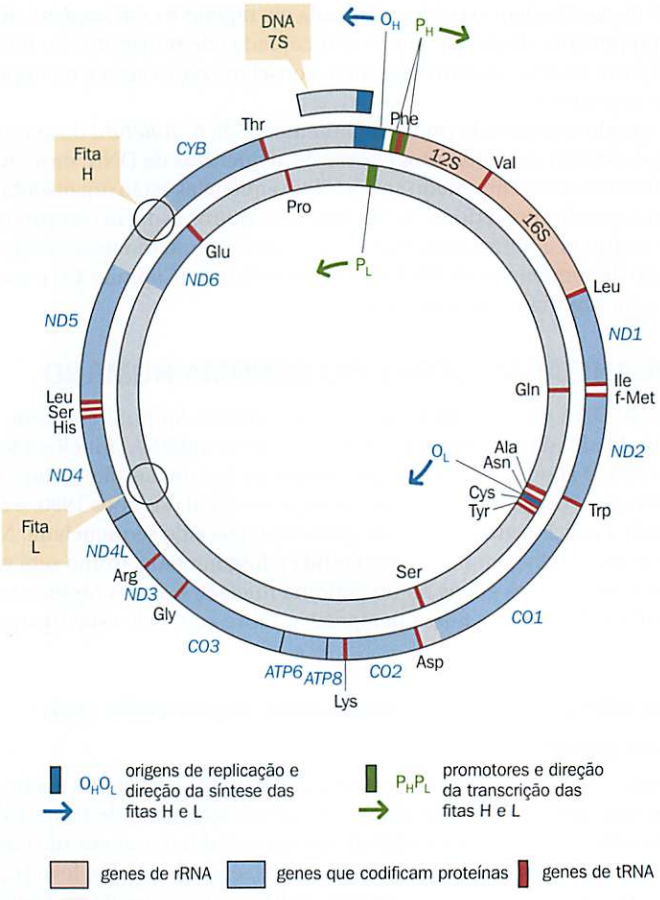


Figura 9.3 A organização do genoma mitocondrial humano. A fita H é transcrita a partir de duas regiões promotoras próximas flanqueando o gene do $tRNA^{Phe}$ (agrupados aqui como P_H); a fita L é transcrita a partir do promotor P_L, na direção oposta. Em ambos os casos, grandes transcritos primários são produzidos e clivados para gerar RNAs de genes individuais. Todos os genes são desprovidos de íntrons e estão localizados em regiões muito próximas. Os símbolos para genes que codificam proteínas são mostrados aqui sem o prefixo *MT* – que significa gene mitocondrial. Os genes que codificam as subunidades 6 e 8 da ATP sintetase (*ATP6* e *ATP8*) são parcialmente sobrepostos. Outros genes que codificam polipeptídeos especificam sete subunidades da NADH-desidrogenase (*ND4L* e *ND1-ND6*), três subunidades de citocromo c oxidase (*CO1-CO3*) e citocromo *b* (*CYB*). Os genes de tRNA estão representados com o nome do aminoácido que eles carregam. A fita curta de DNA 7S é produzida pela síntese repetida de um pequeno segmento da fita H (ver Figura 9.2). Para mais informações, consultar o banco de dados MITOMAP em http://www.mitomap.org/.

Genes mitocondriais e sua transcrição

O genoma mitocondrial humano contém 37 genes, dos quais 28 estão codificados na fita H e os outros nove, na fita L (ver Figura 9.3). Enquanto genes nucleares geralmente possuem seu próprio promotor, a transcrição dos genes mitocondriais lembra aquela dos genes bacterianos. A transcrição do mtDNA começa a partir de promotores comuns na região CR/alça-D e continua ao longo do círculo (em direções opostas para as duas diferentes fitas), gerando grandes transcritos multigênicos. Os RNAs maduros são posteriormente gerados por clivagem dos transcritos multigênicos.

Quase dois terços (24 de 37) dos genes mitocondriais especificam uma molécula de RNA não codificante, funcional, como seu produto final. Existem 22 genes de tRNAs, um para cada um dos 22 tipos de tRNAs mitocondriais. Além disso, dois genes de rRNA são destinados à síntese dos rRNAs 16S e 12S (componentes das subunidades ribossomais grande e pequena, respectivamente). Os 13 genes restantes codificam polipeptídeos, os quais são sintetizados nos ribossomos mitocondriais. Estes 13 polipeptídeos formam parte dos complexos respiratórios mitocondriais, as enzimas de fosforilação oxidativa envolvidas na produção de ATP. Entretanto, a grande maioria dos polipeptídeos que constituem o sistema de fosforilação oxidativa mitocondrial, além de todas as outras proteínas mitocondriais, é codificada por genes nucleares (**Tabela 9.1**). Essas proteínas são traduzidas nos ribossomos citoplasmáticos antes de serem importadas para as mitocôndrias.

Diferente de seu equivalente nuclear, o genoma mitocondrial humano é extremamente compacto: nenhum dos 37 genes mitocondriais possui íntrons, e eles estão bastante agrupados (em média, há um gene por 0,45 kb). As sequências codificantes de alguns genes (notadamente aqueles que codificam a sexta e a oitava subunidades da ATP sintetase mitocondrial) apresentam alguma sobreposição (**Figura 9.4**), e, na maioria dos outros casos, as sequências codificantes de genes vizinhos são contíguas ou separadas por uma ou duas bases não codificantes. Alguns genes não possuem nem mesmo códons de terminação; para superar essa deficiência, códons UAA precisam ser introduzidos no nível pós-transcricional (ver Figura 9.4).

TABELA 9.1 A autonomia limitada do genoma mitocondrial

Componente mitocondrial	Codificado por	
	Genoma mitocondrial	**Genoma nuclear**
Componentes do sistema de fosforilação oxidativa	**13 subunidades**	**80 subunidades**
I NADH desidrogenase	7	42
II Sucinato CoQ redutase	0	4
III Complexo citocromo b-c_1	1	10
IVComplexo citocromo c oxidase	3	10
V Complexo ATP sintetase	2	14
Componentes do aparato de síntese proteica	**24 RNAs**	**79 proteínas**
rRNA	2	0
tRNA	22	0
Proteínas ribossomais	0	79
Outras proteínas mitocondriais	**0**	**All**[a]

[a] Inclui DNA e RNA-polimerases mitocondriais, além de várias outras enzimas, proteínas estruturais e de transporte, etc.

O código genético mitocondrial

Os genomas procariótico e nuclear dos eucariotos codificam muitas centenas e geralmente muitos milhares de proteínas diferentes. Eles estão sujeitos a um código genético universal que é mantido conservado: mutações que poderiam potencialmente alterar o código genético geralmente produzem proteínas cujo mau funcionamento é crítico e, portanto, sofrem forte pressão seletiva negativa. Entretanto, o pequeno genoma mitocondrial produz poucos polipeptídeos. Consequentemente, o código genético mitocondrial conseguiu derivar via mutações para um *código ligeiramente* distinto do código genético universal.

No código genético mitocondrial há 60 códons que determinam aminoácidos específicos, um a menos que no código genético nuclear. Há quatro códons de parada: UAA e UAG (os quais também servem como códons de parada no código genético nuclear) e AGA e AGG (que especificam arginina no código genético nuclear; ver Figura 1.25). O códon de parada nuclear UGA codifica triptofano na mitocôndria, e o AUA especifica metionina em vez de isoleucina.

O genoma mitocondrial especifica todas as moléculas de rRNA e tRNA necessárias para sintetizar as proteínas nos ribossomos mitocondriais, mas depende de genes do genoma nuclear para fornecer os demais componentes, tais como as proteínas que compõem o ribossomo mitocondrial e as aminoacil-tRNA sintetases. Como há apenas 22 tipos de tRNAs mitocondriais humanos, moléculas de tRNA individuais precisam interpretar diferentes códons. Isto é possível devido à *oscilação da terceira base* na interpretação dos códons. Oito das 22 moléculas de tRNA possuem anticódons que reconhecem, cada um, famílias de quatro códons que diferem apenas na terceira base. Os outros 14 tRNAs reconhecem pares de códons que são idênticos nas duas primeiras bases e compartilham ou uma purina ou uma pirimidina na terceira base. Assim sendo, os 22 tRNAs mitocondriais são capazes de reconhecer um total de 60 códons [$(8 \times 4) + (14 \times 2)$].

Além das diferenças nas suas capacidades genéticas e nos seus códigos genéticos, os genomas mitocondrial e nuclear diferem em vários outros aspectos de sua organização e expressão (**Tabela 9.2**).

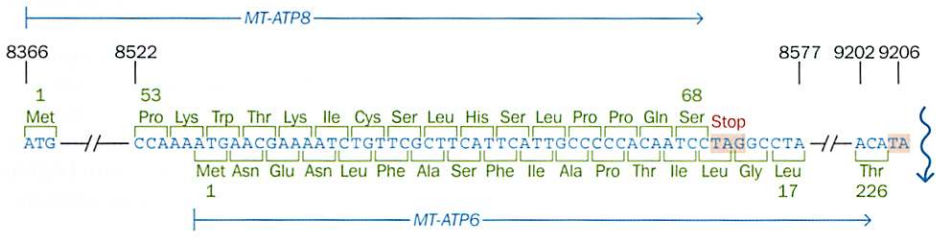


Figura 9.4 Os genes *MT-ATP6* e *MT-ATP8* são transcritos em diferentes quadros de leitura a partir de segmentos sobrepostos na fita H do DNA mitocondrial. *MT-ATP8* é transcrito dos nucleotídeos 8366 ao 8569, e *MT-ATP6*, de 8527 a 9204. Após a transcrição, o RNA que codifica a subunidade da ATP sintetase 6 é clivado depois da posição 9206 e poliadenilado, resultando em um códon UAA C-terminal onde os dois primeiros nucleotídeos são derivados do TA nas posições 9205-9206 e o terceiro nucleotídeo é o primeiro A da cauda de poli(A).

TABELA 9.2 Os genomas nuclear e mitocondrial humanos

	Genoma nuclear	Genoma mitocondrial
Tamanho	3,1 Gb	16,6 kb
Número de diferentes moléculas de DNA	23 (em células XX) ou 24 (em células XY); todas lineares	Uma molécula de DNA circular
Número total de moléculas de DNA por célula	Varia de acordo com a ploidia; 46 em células diploides	Geralmente, vários milhares de cópias (mas o número de cópias varia em diferentes tipos celulares)
Proteínas associadas	Várias classes de histonas e proteínas não histônicas	Amplamente livre de proteínas
Número de genes que codificam proteínas	~21.000	13
Número de genes de RNA	Incerto, mas > 6.000	24
Densidade gênica	~1/120 kb, mas há grande incerteza	1/0,45 kb
DNA repetitivo	Mais de 50% do genoma; ver Figura 9.1	Muito pouco
Transcrição	Genes geralmente transcritos de maneira independente	Transcritos multigênicos são produzidos de ambas as fitas, pesada (H) e leve (L)
Íntrons	Encontrados na maioria dos genes	Ausentes
Porcentagem de DNA codificante de proteínas	~1,1%	~66%
Uso de códons	61 códons de aminoácidos e 3 códons de parada[a]	60 códons de aminoácidos e 4 códons de parada[a]
Recombinação	Pelo menos um evento para cada par de cromossomos homólogos na meiose	Não evidente
Herança	Mendeliana para cromossomo X e autossomos; paterna para cromossomo Y	Exclusivamente materna

[a] Para mais detalhes, ver Figura 1.25.

O genoma nuclear humano consiste em 24 moléculas de DNA cromossômico amplamente diferentes

O genoma nuclear humano possui um tamanho de 3,1 Gb (3.100 Mb) e está distribuído em 24 tipos diferentes de moléculas lineares de DNA dupla-fita, cada uma das quais ligada a proteínas histonas e não histonas, constituindo um cromossomo. Há 22 tipos de autossomos e dois cromossomos sexuais, X e Y. Os cromossomos humanos podem ser facilmente diferenciados pelo bandeamento cromossômico (ver Figura 2.15) e foram classificados em grupos de acordo com seu tamanho e, até certo ponto, de acordo com a posição do centrômero (ver Tabela 2.3).

Existe apenas uma cópia do genoma nuclear em espermatozoides e óvulos, e apenas duas cópias na maioria das células somáticas, em contraste às centenas ou mesmo milhares de cópias do genoma mitocondrial. No entanto, como o tamanho do genoma nuclear é aproximadamente 186 mil vezes maior do que o da molécula de mtDNA, o núcleo de uma célula humana contém mais de 99% do DNA total da célula; o oócito é uma exceção notável, uma vez que contém até 100 mil moléculas de mtDNA.

A sequência total do genoma humano ainda não foi determinada. O Projeto Genoma Humano estava focado principalmente no sequenciamento da *eucromatina*, regiões transcricionalmente ativas, ricas em genes, do genoma nuclear e que correspondem a cerca de 2,9 Gb. Os 200 Mb restantes são constituídos por heterocromatina permanentemente condensada e transcricionalmente inativa (constitutivas). A heterocromatina é constituída por longos arranjos de DNA altamente repetitivos que são difíceis de se sequenciarem corretamente. Por razões semelhantes, os longos arranjos de unidades de transcrição repetidas em *tandem* que codificam os rRNAs 28S, 18S e 5,8S também não foram sequenciados.

O DNA dos cromossomos humanos varia consideravelmente em tamanho e também nas proporções de eucromatina e heterocromatina constitutiva (**Tabela 9.3**). Cada cromossomo possui determinada quantidade de heterocromatina constitutiva no centrômero. Alguns, notadamente os cromossomos 1, 9, 16 e 19, também apresentam quantidades significativas de heterocromatina na região eucromática próxima ao centrômero (*região pericentromérica*), e os cromossomos acrocêntricos possuem duas regiões heterocromáti-

TABELA 9.3 Conteúdo de DNA dos cromossomos humanos

Cromossomo	DNA total (Mb)	Eucromatina (Mb)	Heterocromatina (Mb)	Cromossomo	DNA total (Mb)	Eucromatina (Mb)	Heterocromatina (Mb)
1	249	224	19,5	13	115	96,3	17,2
2	243	240	2,9	14	107	88,3	17,2
3	198	197	1,5	15	103	82,1	18,3
4	191	188	3	16	90	79	10
5	181	178	0,3	17	81	78,7	7,5
6	171	168	2,3	18	78	74,6	1,4
7	159	156	4,6	19	59	60,8	0,3
8	146	143	2,2	20	63	60,6	1,8
9	141	120	18	21	48	34,2	11,6
10	136	133	2,5	22	51	35,1	14,3
11	135	131	4,8	X	155	151	3
12	134	131	4,3	Y	59	26,4	31,6

Os tamanhos dos cromossomos foram retirados do ENSEMBL Human Map View (http://www.ensembl.org/Homo_sapiens/Location/Genome). Os números referentes à heterocromatina são estimativas retiradas do International Genome Sequencing Consortium (2004) *Nature* 431, 931–945. O tamanho do genoma humano total é estimado em 3,1 Gb, com a eucromatina correspondendo a cerca de 2,9 Gb e a heterocromatina a 200 Mb.

cas consideráveis. Porém, a representação mais significativa está no cromossomo Y, onde a maior parte do DNA está organizada sob a forma de heterocromatina.

A composição de bases do componente eucromático do genoma humano é, em média, de 41% (G+C), mas há variação considerável entre os diferentes cromossomos, de 38% de G+C para os cromossomos 4 e 13 até 49% no cromossomo 19. Esta composição também varia muito ao longo dos cromossomos. Por exemplo, o conteúdo médio (G+C) no cromossomo 17q é de 50% na região distal de 10,3 Mb, mas cai para 38% na região adjacente de 3,9 Mb. Existem regiões com menos de 300 kb com variações ainda mais amplas, por exemplo, de 33,1% para 59,3% (G+C).

A proporção de algumas combinações de nucleotídeos pode variar consideravelmente. A exemplo de outros genomas nucleares de vertebrados, o genoma nuclear humano apresenta uma reduzida quantidade de dinucleotídeos CpG. No entanto, algumas regiões pequenas e transcricionalmente ativas do DNA possuem a densidade esperada de CpG e, significativamente, encontram-se não metiladas ou hipometiladas (*ilhas CpG*; **Quadro 9.1**).

O genoma humano contém pelo menos 26 mil genes, mas é difícil determinar seu número exato

Muitos anos depois do Projeto Genoma Humano ter produzido a primeira referência para a sequência do genoma humano, ainda há considerável incerteza acerca do número total de genes em humanos. Quando as primeiras análises do genoma foram publicadas em 2001, o catálogo de genes gerado pelo Consórcio Internacional de Sequenciamento do Genoma Humano era bastante orientado para aqueles que codificam proteínas. As estimativas originais sugeriam a existência de mais de 30 mil genes humanos codificantes de proteínas, dos quais a maioria era referente a predições gênicas sem qualquer evidência experimental. Esse número foi uma superestimativa devido a erros na definição dos genes (ver Quadro 8.5).

Para validar as predições gênicas, evidências que as confirmassem foram buscadas principalmente por comparações evolutivas. As comparações com outros genomas de mamíferos, como o do camundongo e o canino, foram incapazes de identificar genes equivalentes para muitos dos genes humanos originalmente propostos. Ao final de 2009, o número estimado de genes humanos codificantes parecia ter se estabilizado em torno de 20 a 21mil, porém grandes incertezas permaneceram acerca do número de genes humanos de RNA. Genes de RNA são difíceis de identificar por meio de programas de computador que analisam sequências gênicas: não há quadros de leitura aberta (*open reading frames*) para serem identificados, e diversos genes de RNA são muito pequenos e geralmente não muito conservados ao longo da evolução. Existe ainda o problema de como definir um gene de RNA. Conforme detalhado no Capítulo 12, análises abrangentes sugeriram

QUADRO 9.1 Metilação do DNA animal e ilhas CpG de vertebrados

A metilação do DNA em animais pluricelulares geralmente envolve a metilação de uma proporção dos resíduos de citosina, produzindo 5-metilcitosina (mC). Na maioria dos animais (mas não em *Drosophila melanogaster*), o dinucleotídeo CpG é um alvo comum para a metilação de citosinas por citosina-metiltransferases específicas, formando mCpG (**Figura 1A**).

A metilação do DNA tem consequências importantes para a expressão gênica e permite que padrões característicos de expressão de genes sejam estavelmente transmitidos para células-filhas. Este processo também foi implicado em sistemas de defesa do hospedeiro contra transposons. Os vertebrados possuem os mais altos níveis de 5-metilcitosina do reino animal, e a metilação é dispersada por meio dos genomas vertebrados. Entretanto, apenas uma pequena porcentagem das citosinas é metilada (cerca de 3% no DNA humano, a maior parte delas como mCpG, mas com uma pequena porcentagem de mCpNpG, em que N representa um nucleotídeo qualquer).

A 5-metilcitosina é quimicamente instável e tende a sofrer desaminação (ver Figura 1A). Outras bases desaminadas produzem derivados que são identificados como anormais, sendo removidos pela maquinaria de reparo de DNA (por exemplo, citosinas não metiladas produzem uracilas quando desaminadas). Entretanto, a 5-metilcitosina produz timina quando desaminada, uma base natural no DNA que não será reconhecida como anormal pelos sistemas celulares de reparo do DNA. Ao longo de períodos evolutivos, portanto, o número de dinucleotídeos CpG no DNA dos vertebrados foi sendo gradualmente reduzido por conta da conversão lenta, porém constante, de **C**pG em **T**pG (e Cp**A** na fita complementar; ver Figura 1B).

Embora a frequência total de CpG no genoma dos vertebrados seja baixa, há pequenas regiões de DNA não metilado ou hipometilado que são caracterizadas pela presença de frequências *normais* esperadas de CpG. Tais ilhas com densidade normal de CpG (**ilhas CpG**) são comparativamente ricas em GC (contendo, em geral, mais de 50% de GC) e se estendem por centenas de nucleotídeos. Ilhas CpG são consideradas marcadores gênicos porque estão associadas com regiões transcricionalmente ativas. Regiões de DNA altamente metiladas tendem a adotar uma conformação de cromatina condensada, mas em regiões de DNA ativamente transcritas, a cromatina necessita estar em uma conformação não metilada mais aberta, que permita a ligação de várias proteínas regulatórias aos promotores e a outras regiões que controlam a expressão gênica.

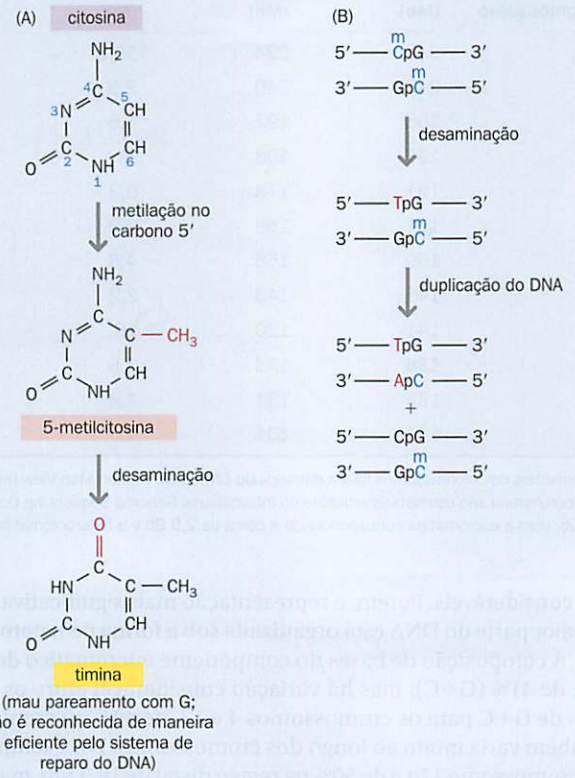


Figura 1 Instabilidade dos dinucleotídeos CpG em vertebrados. (A) A citosina em dinucleotídeos CpG é um alvo para a metilação no átomo de carbono 5′. A 5-metilcitosina resultante é desaminada, produzindo timina (T), a qual não será reconhecida pelos sistemas de reparo de DNA e então tende a ser mantida (entretanto, a desaminação de citosinas não metiladas produz uracila, a qual é facilmente reconhecida pelos sistemas de reparo do DNA). (B) O dinucleotídeo CpG em vertebrados está sendo gradualmente substituído por TpG e CpA.

recentemente que a grande maioria do genoma – e, provavelmente, pelo menos 85% dos nucleotídeos – é transcrita. Atualmente, não se sabe quanto da atividade transcricional se refere à interferência e quanto é funcionalmente significativa.

Em meados de 2009, foram obtidas evidências para, pelo menos, 6 mil genes humanos de RNA, incluindo milhares de genes codificando longas moléculas de RNA não codificante que se acredita serem importantes na regulação da expressão gênica. Além disso, há evidências para dezenas de milhares de pequenos RNAs humanos diferentes, mas em muitos desses casos, um grande número de RNAs pequenos diferentes é obtido a partir do processamento de um único transcrito de RNA. Os RNAs não codificantes serão detalhados na Seção 9.3.

A combinação de cerca de 20 mil genes que codificam proteínas e pelo menos 6 mil genes de RNA alcança um número de, ao menos, 26 mil genes humanos. Este valor continua sendo um número total de genes provisório; definir os genes de RNA é um desafio, e levará algum tempo antes que se possa chegar ao número exato de genes humanos.

Os genes humanos são distribuídos de maneira desigual tanto entre como dentro dos cromossomos

Os genes humanos estão distribuídos de maneira desigual nas moléculas de DNA nuclear. As regiões de heterocromatina constitutiva são desprovidas de genes e, mesmo na porção eucromática do genoma, a densidade gênica pode variar substancialmente entre regiões cromossômicas e também entre cada um dos cromossomos.

A primeira visão geral da distribuição dos genes pelo genoma humano foi obtida quando frações purificadas de ilhas CpG foram hibridizadas com cromossomos em metáfase. Sabe-se, há muito tempo, que ilhas CpG estão fortemente associadas aos genes (ver Quadro 9.1). Com base nisso, concluiu-se que a densidade gênica deveria ser maior em regiões subteloméricas e que alguns cromossomos (p. ex., os cromossomos 19 e 22) são ricos em genes, enquanto outros (p. ex., X e 18) são pobres em genes (ver Figura 8.17). As predições acerca da densidade diferencial de ilhas CpG e de genes foram subsequentemente confirmadas pela análise da sequência do genoma humano.

Esta diferença na densidade gênica também pode ser observada pela coloração Giemsa de cromossomos (bandeamento G). Regiões com baixo conteúdo (G+C) correlacionam-se com as bandas G mais escuras; aquelas com alto conteúdo (G+C), com as bandas claras. Cromossomos e regiões ricos em GC (p. ex., o cromossomo 19 e bandas G claras) são também, comparativamente, ricos em genes. Por exemplo, o complexo de antígenos leucocitários humanos (HLA), rico em genes (180 genes codificantes de proteínas em uma região de 4 Mb), está localizado na banda G clara 6p21.3. Em nítido contraste, o gene da distrofina dos mamutes se estende por 2,4 Mb de DNA em uma banda G escura no Xp21.2, sem evidências da existência de qualquer outro gene codificante de proteínas nesta região.

A duplicação de segmentos de DNA resultou em variação no número de cópias e em famílias gênicas

Genomas pequenos, como os de bactérias e o das mitocôndrias, são geralmente organizados com informação genética densamente agrupada e apresentada de forma extremamente econômica. Genomas grandes, como o genoma nuclear dos eucariotos, e especialmente o genoma dos vertebrados, podem não apresentar essas limitações. O DNA repetitivo é uma característica marcante de grandes genomas, tanto em abundância como em importância.

Tipos diferentes de sequências de DNA podem ser repetidos. Algumas são sequências pequenas não codificantes que estão presentes de poucas a milhões de cópias. Estas serão discutidas mais adiante, na Seção 9.4. Muitas outras são sequências de DNA de tamanho moderadamente longo a grande que geralmente contêm genes ou porções de genes. Estas sequências duplicadas são sujeitas a vários mecanismos genéticos que resultam em *variação no número de cópias* (*CNV*), na qual o número de cópias de sequências específicas, moderadamente longas – geralmente de muitas quilobases a várias megabases de comprimento – varia entre diferentes haplótipos. A variação no número de cópias leva a um tipo de *variação estrutural* que será considerada em mais detalhes no Capítulo 13, mas serão considerados a seguir alguns mecanismos pelos quais os genes são duplicados. Está claro, no entanto, que a CNV é bastante extensa no genoma humano. Por exemplo, quando o genoma de James Watson foi sequenciado, 1,4% dos dados totais obtidos no sequenciamento não mapearam com a sequência de referência do genoma humano. À medida que o sequenciamento do genoma de indivíduos acelera, novas regiões CNV estão sendo identificadas com implicações importantes para a expressão gênica e as doenças.

A duplicação repetida de sequências que contêm genes dá origem a **famílias gênicas**. Conforme será visto nas Seções 9.2 e 9.3, muitos genes humanos são membros de famílias multigênicas que podem variar enormemente em termos de número de cópias e distribuição. Elas surgem por um ou mais mecanismos diferentes que resultam em duplicação gênica. Famílias gênicas podem também conter sequências evolutivamente relacionadas que não atuam mais como genes (*pseudogenes*).

Mecanismos de duplicação gênica

A duplicação gênica foi um evento comum na evolução dos grandes genomas nucleares encontrados em eucariotos complexos. As famílias multigênicas resultantes possuem de duas a muitas cópias gênicas. As cópias gênicas podem estar agrupadas em uma localização subcromossômica ou podem estar dispersas por várias localizações cromossomais. Vários tipos diferentes de duplicação gênica podem ocorrer:

- *Duplicações gênicas em tandem* geralmente surgem pelo *crossing-over* entre cromátides mal-alinhadas, em cromossomos homólogos (*crossing-over desigual*) ou no mesmo cromossomo (*troca desigual entre cromátides-irmãs*). A **Figura 9.5** ilustra o mecanismo descrito de modo geral. O segmento repetido pode ter um tamanho redu-

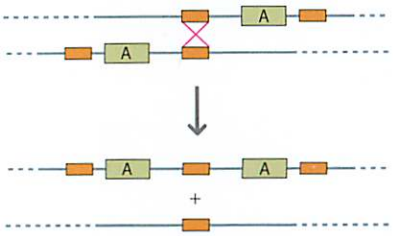


Figura 9.5 Duplicação gênica em *tandem*. O *crossing-over* entre cromátides inadequadamente alinhadas pode resultar em um cromossomo com uma duplicação em *tandem* de uma sequência contendo um gene (como o gene A, representado aqui por um retângulo verde) e em outro cromossomo no qual o gene é perdido. O mau pareamento das cromátides pode ser estabilizado por membros relacionados de famílias de sequências de DNA intercalante repetitivo, como as repetições Alu (conforme demonstrado aqui por retângulos laranjas). O evento de *crossing-over* que leva à duplicação gênica em *tandem* pode resultar de *crossing-over* desigual (*crossing-over* entre cromátides desalinhadas em cromossomos homólogos) ou de troca desigual entre cromátides-irmãs (o processo análogo pelo qual cromátides-irmãs estão desalinhadas; ver Figura 13.3 para uma ilustração).

zido de algumas quilobases ou ser relativamente grande, contendo um ou mais genes. Duas repetições deste tipo são chamadas de *repetições diretas* se o início de uma delas está unido ao final de sua vizinha (→→), ou de *repetições invertidas* quando há junção das regiões de início (→←) ou do final de cada uma delas (←→). Em uma escala de tempo longa (evolutiva), as sequências duplicadas podem ser separadas sobre o mesmo cromossomo (por inserção ou inversão do DNA) ou ser distribuídas em diferentes cromossomos por translocações.

- A *transposição duplicativa* consiste no processo pelo qual uma cópia de DNA duplicado se integra em uma nova localização subcromossômica. Isso geralmente envolve a *retrotransposição*: transcriptases reversas celulares sintetizam uma cópia de cDNA a partir de um transcrito de RNA, e esta cópia de cDNA se integra em uma nova localização cromossômica. O mesmo tipo de mecanismo pode gerar cópias defeituosas de genes, o que será detalhado na Seção 9.2.
- *Duplicação gênica por fusão de células ancestrais.* Acredita-se que células eucarióticas aeróbias tenham se desenvolvido devido à endocitose de um tipo de célula bacteriana por uma célula eucariótica precursora. Acredita-se que o genoma mitocondrial atual tenha se originado a partir do genoma bacteriano, do qual seria apenas um pequeno remanescente, já que muitos dos genes bacterianos originais foram subsequentemente excisados e transferidos para o genoma nuclear atual. Consequentemente, o genoma nuclear contém genes duplicados que codificam isoformas de certas enzimas e de outras proteínas-chave que são específicas e direcionadas para o citoplasma e para a mitocôndria.
- *Duplicações subgenômicas de grande escala* podem surgir em consequência de translocações cromossômicas. Regiões eucromáticas próximas aos centrômeros e telômeros humanos (regiões pericentroméricas e subteloméricas, respectivamente) são relativamente instáveis e sujeitas à recombinação com outros cromossomos. Consequentemente, grandes segmentos de DNA contendo múltiplos genes foram duplicados. Ao longo dos últimos 40 milhões de anos da evolução dos primatas, este processo levou à duplicação de aproximadamente 400 grandes segmentos de DNA (com várias megabases de comprimento), correspondentes a mais de 5% do genoma eucromático. Este tipo de duplicação, conhecido como **duplicação segmentar**, resulta em uma identidade de sequência extremamente alta (geralmente maior do que 95%) entre as cópias de DNA e pode envolver tanto duplicações intracromossômicas como intercromossômicas (**Figura 9.6**). Duplicações segmentares contribuem significativamente para a variação do número de cópias e para rearranjos cromossômicos que levam a doenças e à rápida inovação gênica. Sua origem será considerada no Capítulo 10.
- *Duplicação do genoma inteiro.* Sabe-se hoje, a partir de estudos de genômica comparada, que duplicações do genoma inteiro ocorreram diversas vezes ao longo da evolução, em uma ampla variedade de linhagens eucarióticas. Por exemplo, há evidências convincentes de que uma duplicação do genoma inteiro teria acontecido no início da evolução dos cordados. Esse tipo de evento poderia explicar por que os vertebrados possuem quatro agrupamentos HOX (ver Figura 5.5). A duplicação do genoma inteiro é detalhada no Capítulo 10, no contexto da evolução do genoma.

9.2 GENES QUE CODIFICAM PROTEÍNAS

Durante muitos anos, os geneticistas moleculares acreditaram que o principal produto funcional do DNA era a proteína. Estudos de genomas procarióticos apoiavam esta crença,

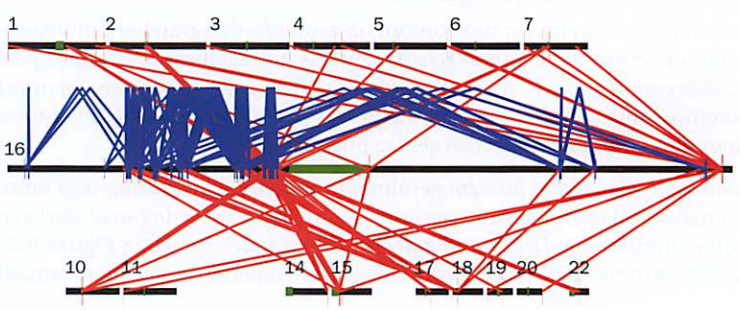


Figura 9.6 Duplicação segmentar. A barra horizontal no centro da figura é um mapa linear do DNA do cromossomo humano 16 (o segmento verde central representa a heterocromatina). As barras pretas horizontais no topo e na base representam mapas lineares de 16 outros cromossomos que contêm grandes segmentos compartilhados com o cromossomo 16, com linhas vermelhas conectoras marcando a posição das sequências homólogas. Duplicações intracromossômicas são representadas por bifurcações azuis (^) ligando as posições de grandes sequências duplicadas no cromossomo 16. [Reproduzido de Martin J (...). Com permissão de Macmillan Publishers Ltd. Para coordenadas cromossômicas e mais informações sobre duplicação segmentar no genoma humano, bancos de dados dedicados à duplicação segmentar podem ser acessados em http://humanparalogy.gs.washington.edu/ e http://projects.tcag.ca/humandup/.]

em parte porque estes genomas são ricos em DNA que codifica proteínas. Portanto, foi uma surpresa descobrir que os genomas muito maiores dos eucariotos complexos apresentam, comparativamente, pouco DNA que codifica proteínas. Por exemplo, sequências de DNA que codificam proteína correspondem a quase 90% do genoma da *E. coli*, mas apenas 1,1% do genoma humano.

Genes humanos que codificam proteínas variam muito em tamanho e organização interna

Variação de tamanho

Os genes de organismos simples, tais como bactérias, são relativamente semelhantes em tamanho e são geralmente muito pequenos (contendo 1 kb de extensão). Em eucariotos complexos, os genes podem apresentar uma enorme variação de tamanho. Embora geralmente exista uma correlação direta entre o tamanho do gene e de seu produto, há algumas anomalias marcantes. Por exemplo, o gigantesco gene da distrofina, com 2,4 Mb, é mais de 50 vezes maior do que o gene da apolipoproteína B, mas a proteína distrofina tem um comprimento linear (número total de aminoácidos) que é cerca de 80% aquele da apolipoproteína B (**Tabela 9.4**).

Uma pequena minoria dos genes humanos que codificam proteína é desprovida de íntrons e representada por genes geralmente pequenos (ver nota da Tabela 9.4 para alguns exemplos). Para aqueles que possuem íntrons, há uma correlação inversa entre o tamanho do gene e a fração de DNA codificante (ver Tabela 9.4). Isso não ocorre porque éxons de genes grandes são menores do que aqueles de genes menores. O tamanho médio dos éxons em genes humanos é de aproximadamente 300 pb, e o tamanho do éxon é comparativamente independente do comprimento do gene. Contrariamente, há grande variação no comprimento dos íntrons, e genes grandes tendem a apresentar íntrons bastante grandes (ver Tabela 9.4). No entanto, a transcrição de íntrons grandes é custosa em termos de tempo e energia; a transcrição das 2,4 Mb do gene da distrofina leva aproximadamente 16 horas. Portanto, genes muito expressos geralmente possuem íntrons pequenos ou não os possuem.

TABELA 9.4 Variação estrutural em tamanho e organização dos genes humanos que codificam proteínas

Proteína humana	Tamanho da proteína (nº de aminoácidos)	Tamanho do gene (kb)	Nº de éxons	DNA codificante (%)	Tamanho médio do éxon (pb)	Tamanho médio do íntron (pb)
SRY	204	0,9	1	94	850	—
β-globina	146	1,6	3	38	150	490
p16	156	7,4	3	17	406	3.064
Albumina sérica	609	18	14	12	137	1.100
Colágeno tipo VII	2.928	31	118	29	77	190
p53	393	39	10	6	236	3.076
Complemento C3	1.641	41	29	8,6	122	900
Apolipoproteína B	4.563	45	29	31	487	1.103
Fenilalanina hidroxilase	452	90	26	3	96	3.500
Fator VIII	2.351	186	26	3	375	7.100
Huntingtina	3.144	189	67	8	201	2.361
Proteína RB1 do retinoblastoma	928	198	27	2,4	179	6.668
CFTR (receptor transmembrana da fibrose cística)	1.480	250	27	2,4	227	9.100
Titina	34.350	283	363	40	315	466
Utrofina	3.433	567	74	2,2	168	7.464
Distrofina	3.685	2.400	79	0,6	180	30.770

Quando da existência de isoformas, os números fornecidos representam as maiores isoformas. À medida que os genes aumentam de tamanho, o tamanho dos éxons permance relativamente constante, mas o tamanho dos íntrons pode aumentar bastante. Éxons internos tendem a ser uniformes em tamanho, mas éxons terminais ou alguns éxons próximos à extremidade 3′ podem ter tamanho igual a várias quilobases; por exemplo, o éxon 26 do gene *APOB* tem 7,5 kb. Observe o grande conteúdo de éxons e o tamanho comparativamente menor dos íntrons nos genes que codificam o colágeno tipo VIII e a titina. Além do *SRY*, outros genes que codificam proteínas de éxon único no genoma nuclear incluem retrogenes (ver Tabela 9.8) e genes codificando outras proteínas SOX, interferons, histonas, diversos receptores acoplados a proteínas G, proteínas de choque térmico, muitas ribonucleases e vários receptores de neurotransmissores e receptores de hormônios.

Sequências repetitivas no interior do DNA codificante

Sequências de DNA altamente repetitivas são muitas vezes encontradas no interior de íntrons e em sequências flanqueadoras dos genes. Elas serão detalhadas na Seção 9.4. Além disso, sequências de DNA repetitivo são encontradas nos éxons, em diferentes graus. Repetições em *tandem* de pequenas sequências de oligonucleotídeos (1-4 pb) são frequentes e podem simplesmente ser o reflexo das frequências estatisticamente esperadas para certas composições de bases. Repetições de sequências em *tandem* que codificam domínios proteicos conhecidos ou supostos também são comuns e podem ser funcionalmente vantajosas, pois fornecem alvos biológicos mais disponíveis.

A identidade de sequência entre domínios proteicos repetidos é geralmente baixa, mas às vezes pode ser alta. A lipoproteína Lp (a), codificada pelo gene *LPA* no cromossomo 6q26, fornece um exemplo clássico. Ela possui múltiplos domínios do tipo *kringle* repetidos em *tandem*, cada um deles com um tamanho aproximado de 114 aminoácidos, formando grandes alças unidas por pontes dissulfeto. Os diferentes domínios *kringle* são geralmente quase idênticos em sequência de aminoácidos. Mesmo em nível de sequência de nucleotídeos, as repetições de DNA que codificam os domínios *kringle* apresentam alto grau de identidade de sequência, tornando-os sujeitos a *crossing-over* desigual. Consequentemente, o gene *LPA* está sujeito a polimorfismo de tamanho, e o número de domínios *kringle* na lipoproteína Lp (a) é variável, sendo geralmente igual a 15 ou mais.

Proteínas diferentes podem ser especificadas por unidades de transcrição sobrepostas

Genes sobrepostos e genes dentro de genes

Genomas simples apresentam alta densidade gênica (cerca de um gene a cada 0,5, 1 e 2 kb para os genomas mitocondrial humano, *Escherichia coli* e *Saccharomyces cerevisiae*, respectivamente) e geralmente apresentam exemplos de genes parcialmente sobrepostos. Diferentes quadros de leitura podem ser utilizados, às vezes na mesma fita senso. Em organismos complexos, como os seres humanos, os genes são muito maiores, e há um menor agrupamento das sequências codificadoras de proteínas (**Tabela 9.5**).

A densidade gênica varia enormemente de cromossomo para cromossomo e entre regiões distintas de um mesmo cromossomo. Em regiões cromossômicas com alta densidade gênica, podem-se encontrar genes sobrepostos; eles são geralmente transcritos a partir de fitas opostas de DNA. Por exemplo, a região de classe III do complexo HLA, no cromossomo 6p21.3, tem uma densidade gênica média de cerca de um gene a cada 15 kb e é conhecida por conter diversos exemplos de genes parcialmente sobrepostos (**Figura 9.7A**).

Análises de genoma inteiro mostraram que cerca de 9% dos genes humanos que codificam proteínas se sobrepõem a outros genes codificantes. Mais de 90% das sobreposições envolvem genes transcritos a partir de fitas opostas. Às vezes as sobreposições são parciais, mas em outros casos pequenos genes que codificam proteínas estão localizados no interior de íntrons de genes maiores. O gene *NF1* (neurofibromatose tipo I), p. ex., possui três pequenos genes internos transcritos a partir da fita oposta (**Figura 9.7B**).

Análises recentes também mostraram que genes de RNA muitas vezes podem estar sobrepostos a genes que codificam proteínas. O posicionamento de genes de RNA será abordado na Seção 9.3.

Genes transcritos ou cotranscritos de maneira divergente a partir de um promotor comum

Alguns genes que codificam proteínas compartilham um promotor. Em muitos casos, as extremidades 5′ dos dois genes são separadas por poucas centenas de nucleotídeos, e os genes são transcritos em direções opostas a partir do promotor comum. Este tipo de organização gênica bidirecional pode favorecer a regulação conjunta dos dois genes.

Alternativamente, genes com um promotor comum são transcritos na mesma direção, produzindo transcritos multigênicos que serão então clivados para dar origem a transcritos individuais para cada gene. Estes genes formam uma unidade de transcrição denominada *policistrônica* (= multigênica). Unidades policistrônicas de transcrição são comuns em genomas simples como os de bactérias e mitocôndrias (ver Figura 9.3). No genoma nuclear, alguns exemplos de proteínas diferentes produzidas a partir de uma unida-

TABELA 9.5 Estatísticas do genoma e dos genes humanos

TAMANHO DOS COMPONENTES DO GENOMA	
Genoma mitocondrial	16,6 kb
Genoma nuclear	3,1 Gb[a]
Componente eucromático	2,9 Gb (~93%)
Fração altamente conservada	~150 Mb (~5%)
Sequências de DNA codificante de proteínas	~35 Mb (~1,1%)
Outras sequências de DNA altamente conservadas	~115 Mb (~3,9%)
DNA duplicado segmentalmente	~160 Mb (~5,5%)
DNA altamente repetitivo	~1,6 Gb (~50%)
Heterocromatina constitutiva	~200 Mb (~7%; Tabela 9.3)
Repetições baseadas em transposons	~1,4 Gb (~45%; Tabela 9.12)
DNA por cromossomo	48 Mb–249 Mb (Tabela 9.3)
NÚMERO DE GENES	
Genoma nuclear	> 26.000
Genoma mitocondrial	37
Genes que codificam proteínas	~20.000–21.000
Genes de RNA	> 6.000 (número exato desconhecido)
Pseudogenes relacionados a genes que codificam proteínas	> 12.000
DENSIDADE GÊNICA	
Genoma nuclear	>1 a cada 120 kb (porém, há muita incerteza)
Genoma mitocondrial	1 a cada 0,45 kb
TAMANHO DOS GENES QUE CODIFICAM PROTEÍNAS	
Tamanho médio	53,6 kb
Menor	algumas centenas de pares de base de extensão (vários exemplos)
Maior	2,4 Mb (distrofina)
NÚMERO DE ÉXONS EM GENES QUE CODIFICAM PROTEÍNAS	
Número médio de éxons em um gene[b]	9,8
Maior número em um gene	363 (no gene da titina)
Menor número em um gene	1 (sem íntrons – ver Tabelas 9.4 e 9.7 para exemplos)
TAMANHO DOS ÉXONS EM GENES QUE CODIFICAM PROTEÍNAS	
Tamanho médio do éxon	288 pb (mas éxons na extremidade 3′ dos genes tendem a ser maiores)
Menor	< 10 pb (vários; p. ex., o éxon 3 do gene da troponina I *TNNI1* tem apenas 4 pb)
Maior	18,2 kb (éxon 6 da isoforma 201 de *MUC16*)
TAMANHO DOS ÍNTRONS EM GENES QUE CODIFICAM PROTEÍNAS	
Menor	< 30 pb (vários)
Maior	1,1 Mb (íntron 5 do gene *KCNIP4*)
TAMANHO DO RNA	
Menor RNA não codificante	< 20 nucleotídeos (p. ex., muitos RNAs associados a sítios de início de transcrição têm 18 nucleotídeos)
Maior RNA não codificante	muitas centenas de milhares de nucleotídeos; por exemplo, o RNA antissenso de *UBE3A* tem cerca de 1 Mb
Maior mRNA	> 103 kb (mRNA de titina; isoforma NF-2A)
TAMANHO DO POLIPEPTÍDEO	
Menor	dezenas de aminoácidos (vários neuropeptídeos)
Maior	34.350 aminoácidos (titina, isoforma NF-2A)

[a]Observe que o tamanho total pode variar entre os haplótipos devido à variação no número de cópias. [b]Para a maior isoforma. A maior parte dos dados foi obtida dos 55 conjuntos de dados liberados pelo ENSEMBL.

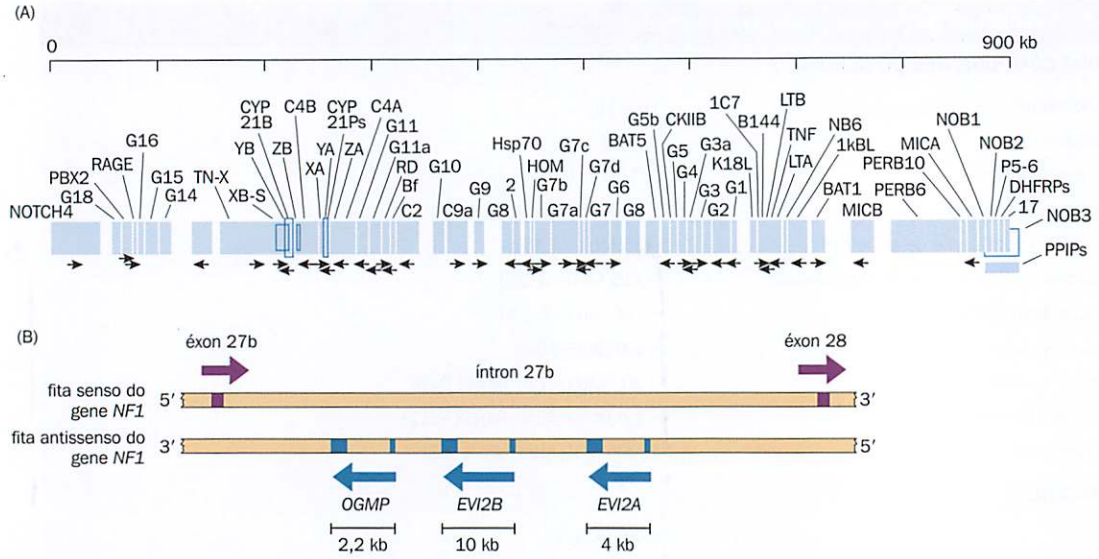


Figura 9.7 Genes sobrepostos e genes dentro de genes. (A) Genes do complexo do HLA de classe III estão altamente agrupados e, em alguns casos, sobrepostos. As setas indicam a direção da transcrição. (B) O íntron 27b do gene *NF1* (neurofibromatose do tipo 1) tem 60,5 kb de comprimento e contém três pequenos genes internos, cada um deles com dois éxons, os quais são transcritos a partir da fita oposta. Os genes internos (não representados em escala) são *OGMP* (glicoproteína mielínica de oligodendrócitos) e *EVI2A* e *EVI2B* (homólogos humanos de genes murinos que parecem estar envolvidos no desenvolvimento de leucemia e que estão localizados em sítios de integração de vírus ecotrópicos).

de comum de transcrição são conhecidos. Geralmente, elas são produzidas por clivagem de uma proteína híbrida precursora que é traduzida a partir de um transcrito comum. As cadeias A e B da insulina, as quais são intimamente relacionadas funcionalmente, são produzidas desta maneira (ver Figura 1.26), bem como os peptídeos hormonais relacionados somatostatina e neuronostanina. Algumas vezes, no entanto, proteínas funcionalmente distintas são produzidas a partir de um precursor proteico comum. Os genes *UBA52* e *UBA80*, por exemplo, geram tanto a ubiquitina como uma proteína ribossomal não relacionada (S27a e L40, respectivamente).

Análises mais recentes mostraram que a tradicional e antiga ideia de que a maioria dos genes humanos são unidades de transcrição independentes não é verdadeira, e, portanto, a definição de gene deverá ser radicalmente revisada. Atualmente, sabe-se que a transcrição multigênica é relativamente frequente no genoma humano, e que proteínas específicas e RNAs não codificantes funcionais podem ser produzidos a partir de RNAs precursores comuns. Este mecanismo será mais explorado na Seção 9.3.

Genes humanos que codificam proteínas frequentemente pertencem a famílias gênicas que podem estar agrupadas ou dispersas em diferentes cromossomos

Genes duplicados e componentes de sequências codificantes duplicados são comuns em genomas animais, considerando-se especialmente os genomas de grandes vertebrados. Conforme será visto no Capítulo 10, a duplicação gênica foi um mecanismo direcionador importante para a evolução da complexidade funcional e a origem de organismos progressivamente mais complexos. Genes que operam na mesma via funcional ou em vias semelhantes, mas que produzem proteínas que apresentam pouca evidência de similaridade de sequência, possuem relação evolutiva distante e geralmente estão dispersos em diferentes localizações cromossômicas. Exemplos incluem os genes que codificam a insulina (no cromossomo 11p) e o receptor de insulina (19p); as cadeias pesada (11q) e leve da ferritina (22q); a esteroide 11-hidroxilase (8q) e a esteroide 21-hidroxilase (6p); e JAK1 (1p) e STAT1 (2q). Entretanto, genes que produzem proteínas estrutural e funcionalmente similares estão frequentemente organizados em agrupamentos gênicos (*gene clusters*).

Classes diferentes de famílias gênicas humanas podem ser reconhecidas pelo grau de similaridade de sequência e estrutura de seus produtos proteicos. Se dois genes diferentes originam produtos proteicos muito semelhantes, é muito provável que tenham sido

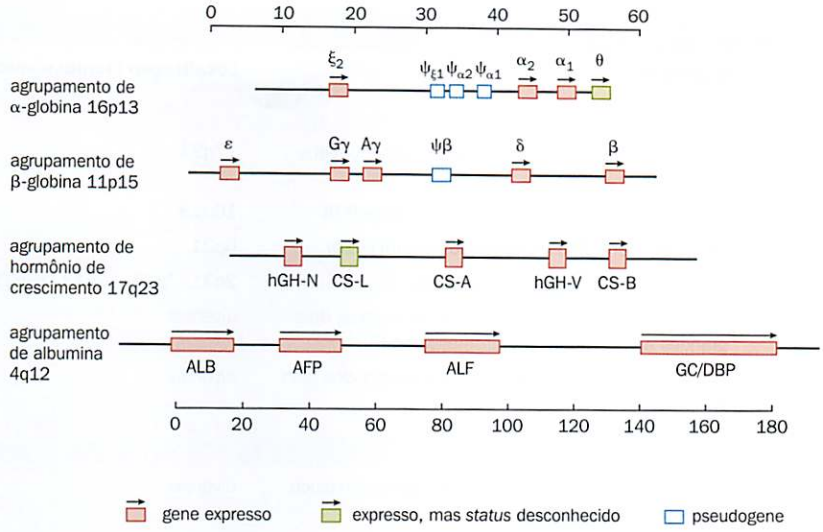


Figura 9.8 Exemplos de famílias de genes humanos agrupados. Genes agrupados são geralmente semelhantes em sequência e são transcritos a partir da mesma fita. Agrupamentos gênicos muitas vezes apresentam uma mistura de genes expressos e pseudogenes não funcionais. O *status* funcional dos genes θ-globina e *CS-L* é desconhecido. As escalas no topo (agrupamentos de globinas e hormônio de crescimento) e na base da figura (agrupamento de albumina) estão em quilobases.

originados por um evento de duplicação gênica evolutivamente recente, provavelmente por algum evento de duplicação gênica em *tandem*, e há uma tendência de que estejam agrupados em uma posição subcromossômica específica. Se estes genes produzem proteínas com sequências menos relacionadas, é mais provável que tenham surgido por um evento de duplicação mais antigo. Originalmente eles poderiam estar agrupados, mas ao longo do tempo, em uma escala de tempo evolutiva, estes genes podem ter sido separados por translocações ou inversões, e há a tendência de que estejam em localizações cromossômicas distintas.

Algumas famílias gênicas estão organizadas em múltiplos agrupamentos. Os genes das globinas β, γ, δ e ε estão localizados em um agrupamento gênico no cromossomo 11p e são mais relacionados entre eles do que com os genes do agrupamento de α-globina no cromossomo 16p (**Figura 9.8**). Os genes de β-globina do agrupamento no cromossomo 11p se originaram por eventos de duplicação gênica evolutivamente muito mais recentes do que os eventos que deram origem aos ancestrais dos genes de α e β-globina. Um exemplo notável de família gênica organizada em agrupamentos múltiplos é o da família gênica de receptores olfativos. Estes genes codificam um repertório diverso de receptores que nos permitem discriminar milhares de odores diferentes; os genes estão localizados em grandes agrupamentos, distribuídos em diversas localizações cromossômicas distintas (**Tabela 9.6**).

Algumas famílias gênicas possuem cópias gênicas individuais localizadas em duas ou mais regiões cromossômicas sem ocorrência de agrupamentos gênicos (ver Tabela 9.6). Os genes em diferentes localizações são geralmente divergentes em sequência, a menos que a duplicação gênica tenha sido relativamente recente ou que tenha havido pressão seletiva considerável para a manutenção da conservação da sequência. Em situações de famílias gênicas dispersas, espera-se que os membros da família tenham se originado a partir de duplicações gênicas antigas.

Classes diferentes de famílias gênicas podem ser reconhecidas pela similaridade de sequência e estrutura de seus produtos proteicos

Conforme listado abaixo, várias classes de famílias gênicas podem ser distinguidas de acordo com o nível de identidade de sequência entre seus genes membros individuais:

- Em famílias gênicas cujos membros apresentam relação próxima, os genes possuem um alto grau de homologia de sequência ao longo da maior parte da sequência codificante. Exemplos incluem as famílias das histonas (as histonas são altamente conservadas, e membros de subfamílias são praticamente idênticos) e as famílias gênicas das α e β-globinas.
- Em famílias gênicas definidas por um domínio proteico comum, os membros podem apresentar homologia de sequência muito baixa, mas possuem certas sequências que

TABELA 9.6 Exemplos de famílias multigênicas agrupadas e dispersas

Família	Nº de cópias	Organização	Localização cromossômica
FAMÍLIAS GÊNICAS AGRUPADAS			
Agrupamento do gene do hormônio do crescimento	5	agrupados em 67 kb de extensão; um pseudogene (Figura 9.8)	17q24
Agrupamento dos genes de α-globina	7	agrupados em ~50 kb de extensão (Figura 9.8)	16p13
Genes da cadeia pesada de HLA de classe I	~20	agrupados em 2 Mb de extensão; (Figura 9.10)	6p21
Genes HOX	38	organizados em quatro agrupamentos (Figura 5.5)	2q31, 7p15, 12q13, 17q21
Família gênica das histonas	61	agrupamentos pequenos em algumas regiões; dois grandes agrupamentos no cromossomo 6	diversas
Família dos genes de receptores olfativos	> 900	cerca de 25 grandes agrupamentos espalhados pelo genoma	diversas
FAMÍLIAS GÊNICAS DISPERSAS			
Aldolase	5	três genes funcionais e dois pseudogenes em cinco cromossomos diferentes	diversas
PAX	9	todos os nove genes são funcionais	diversas
NF1 (neurofibromatose tipo I)	> 12	um gene funcional em 22q11; outras cópias são pseudogenes não processados ou fragmentos gênicos (Figura 9.11)	diversas, a maioria delas pericentromérica
Cadeia pesada da ferritina	20	um gene funcional no cromossomo 11; a maioria das cópias é de pseudogenes processados	diversas

codificam um ou mais domínios proteicos específicos. Exemplos incluem as famílias gênicas PAX e SOX (**Tabela 9.7**).

- Exemplos de famílias gênicas definidas por motivos proteicos pequenos funcionalmente similares incluem famílias de genes que codificam proteínas funcionalmente relacionadas com um motivo *box* DEAD (Asp-Glu-Ala-Asp) ou a repetição WD (**Figura 9.9**).

Alguns genes codificam produtos que são funcionalmente relacionados de uma forma geral, mas apresentam homologia de sequência muito baixa ao longo de um segmento extenso, sem a ocorrência de motivos de aminoácidos significativamente conservados. Apesar disso, pode haver alguma evidência para características estruturais gerais comuns. Tais genes podem ser agrupados em uma **superfamília gênica**, evolutivamente antiga, contendo muitos genes membros. Considerando que múltiplos eventos de duplicação gênica diferentes ocorreram periodicamente durante o longo período evolutivo de uma superfamília gênica, alguns de seus membros produzem proteínas muito divergentes em termos de sequência em relação a outros membros da família, mas um maior grau de relação de sequência será mais facilmente visualizado ente genes resultantes de duplicações mais recentes.

Dois exemplos importantes de superfamílias gênicas são a superfamília das imunoglobulinas (Ig) e a superfamília dos receptores acoplados à proteína G (GPCR). Todos

TABELA 9.7 Exemplos de genes humanos com motivos de sequência que codificam domínios altamente conservados

Família gênica	Número de genes	Sequência motivo/domínio
Genes *homeobox*	38 genes *HOX* mais 197 genes órfãos de *homeobox*	*homeobox* especifica um homeodomínio de ~60 aminoácidos; uma grande variedade de subclasses diferentes foi definida
Genes *PAX*	9	*box* pareado codifica um domínio pareado de ~124 aminoácidos; genes *PAX* geralmente têm um tipo de homeodomínio conhecido como homeodomínio do tipo pareado
Genes *SOX*	19	*box* HMG semelhante a SRY que codifica domínio de 70-80 aminoácidos
Genes *TBX*	14	*box* T codifica um domínio de ~170 aminoácidos
Genes de domínio *forkhead*	50	o domínio tipo *forkhead* possui tamanho de ~110 aminoácidos
Genes de domínio POU	16	o domínio tipo POU possui tamanho de ~150 aminoácidos

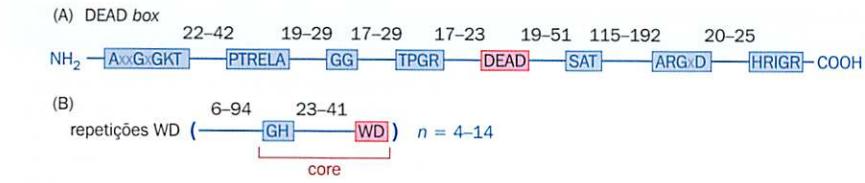


Figura 9.9 Algumas famílias gênicas codificam proteínas funcionalmente relacionadas com motivos curtos de aminoácidos conservados. (A) Motivos da família DEAD *box*. Esta família gênica codifica produtos envolvidos em processos celulares que incluem a alteração da estrutura secundária do RNA como o início da tradução e o *splicing*. Oito motivos de aminoácidos extremamente conservados são evidentes, incluindo o DEAD *box* (Asp-Glu-Ala-Asp). Os números se referem à faixa de tamanho geralmente observado para sequências de aminoácidos intervenientes; X representa um aminoácido qualquer. (B) Motivos da família de repetições WD. Esta família gênica codifica produtos que estão envolvidos em uma ampla variedade de funções regulatórias, tais como a regulação da divisão celular, a transcrição, a sinalização transmembrana e modificações do mRNA. Os produtos gênicos se caracterizam pela presença de 4 a 16 repetições WD em *tandem*, cada uma delas contendo uma sequência central de tamanho fixo que inicia com o dipeptídeo GH (Gly-His) e termina com o dipeptídeo WD (Trp-Asp), precedido por uma sequência de tamanho variável.

os membros da superfamília das Ig possuem domínios globulares semelhantes àqueles encontrados nas imunoglobulinas, e, além destas, a superfamília inclui uma ampla variedade de proteínas de superfície celular e proteínas solúveis envolvidas nos processos de reconhecimento, ligação ou adesão celular (ver Figura 4.22 para alguns exemplos). A superfamília GPCR é muito grande, com pelo menos 799 membros individuais distribuídos pelo genoma humano. Todas as proteínas GPCR possuem uma estrutura comum de sete segmentos transmembrana em α-hélice, mas estes têm, em geral, baixa homologia de sequência (menos de 40%) entre eles. Estes receptores estão envolvidos na sinalização celular induzida por ligantes via interação com proteínas G intracelulares, e a maioria atua como receptores de rodopsina.

Eventos de duplicação gênica que originam famílias multigênicas também criam pseudogenes e fragmentos gênicos

Famílias gênicas muitas vezes apresentam cópias de genes defeituosas além das cópias funcionais. Uma cópia gênica defeituosa que contenha ao menos alguns éxons de um gene funcional é chamada de **pseudogene** (**Quadro 9.2**). Outras cópias gênicas defeituosas podem ter apenas porções limitadas da sequência gênica, às vezes apenas um éxon e, portanto, são algumas vezes descritas como *fragmentos gênicos*.

Famílias gênicas agrupadas geralmente possuem cópias gênicas defeituosas que surgiram por eventos de duplicação em *tandem*. Estas cópias são exemplos de *pseudogenes não processados*. Pode-se detectar que a cópia ocorreu em nível de DNA genômico porque pseudogenes não processados contêm sequências equivalentes a éxons e íntrons e algumas vezes também de regiões promotoras a montante. Entretanto, mesmo que a cópia apresente sequências que correspondem à totalidade do gene funcional, um exame mais detalhado irá revelar a presença de códons de terminação inapropriados em éxons, junções de *splicing* aberrantes e outros. Exemplos clássicos de pseudogenes não processados são encontrados nos agrupamentos das famílias das α e β-globinas (ver Figura 9.8). Às vezes, cópias menores e truncadas de genes e cópias de fragmentos gênicos também são evidentes, como ocorre na família dos genes de HLA de classe I (**Figura 9.10**).

Poucas famílias gênicas que se encontram distribuídas em diferentes regiões cromossômicas também podem apresentar cópias de pseudogenes não processados de um gene funcional determinado. Certos tipos de regiões subcromossômicas, notadamente as regiões pericentroméricas e subteloméricas, são relativamente instáveis. Elas tendem a apresentar eventos de recombinação que podem resultar em segmentos gênicos duplicados (contendo tanto éxons como íntrons) distribuídos para outras regiões cromossômicas. Estas cópias gênicas são geralmente defeituosas porque não possuem algumas sequências do gene funcional. Dois exemplos ilustrativos correspondem a sequências relacionadas aos genes *NF1* (neurofibromatose tipo 1) e *PKD1* (doença policística renal do adulto).

O gene *NF1* está localizado no cromossomo 17q11.2. Devido à instabilidade pericentromérica, várias cópias de fragmentos gênicos/pseudogênicos não processados de *NF1* estão distribuídas ao longo de sete cromossomos diferentes, nove das quais estão localizadas em regiões pericentroméricas (**Figura 9.11A**). O gene *PKD1* possui mais de 46 éxons abrangendo 50 kb e está localizado em uma região subtelomérica do cromossomo 16p13.3. Seis pseudogenes *PKD1* não processados foram gerados por duplicações segmentares ao longo da evolução dos primatas e foram inseridos em uma região que está aproximadamente a 13-16 Mb do gene *PKD1*, correspondendo a uma porção da banda 16p13.1 (**Figura 9.11B**). Estes pseudogenes não possuem sequências da extremidade 3′ do gene *PKD1*, mas contêm sequências equivalentes à maior parte da sequência genômica que vai dos éxons 1–32, apresentando identidade de sequência com *PKD1* entre 97,6% e 97,8%.

QUADRO 9.2 As origens, prevalência e funcionalidade dos pseudogenes

Os pseudogenes são geralmente vistos como cópias defeituosas de um gene funcional com o qual compartilham homologia de sequência significativa. Eles surgem por algum tipo de evento de duplicação gênica que produz duas cópias gênicas. A pressão de seleção para conservar a função do gene precisa ser imposta apenas em uma das cópias gênicas; a outra cópia fica livre para mutar (*deriva genética*) e pode acumular mutações inativantes, originando um pseudogene. Entretanto, algumas sequências são denominadas de pseudogenes mesmo que não tenham sido originadas pela cópia do DNA. Por exemplo, conforme será visto no Capítulo 10, seres humanos possuem *pseudogenes solitários* raros que são claramente ortólogos de genes funcionais de grandes primatas e se tornaram defeituosos após adquirir mutações deletérias na linhagem humana.

Diferentes mecanismos de duplicação gênica podem originar múltiplas cópias gênicas funcionais e pseudogenes defeituosos. A sequência de DNA genômico é copiada, ou uma cópia de cDNA é feita (após a transcrição reversa de um transcrito de RNA processado) e se integra no DNA genômico. No caso de genes que codificam proteínas, a cópia em nível de DNA genômico pode levar à duplicação do promotor e de regiões regulatórias a montante do gene, bem como de éxons e íntrons. Um gene defeituoso derivado de uma cópia de uma sequência de DNA genômico é conhecido como *pseudogene não processado* (**Figura 1A**). Tais pseudogenes geralmente surgem por duplicações em *tandem* e por isso tendem a estar localizados em regiões próximas a seus equivalentes funcionais (ver Figuras 9.8 e 9.10B para exemplos), porém alguns deles são dispersados como resultado de recombinação (ver Figura 9.11 para exemplos).

A cópia em nível de cDNA produz cópias gênicas desprovidas de íntrons, elementos do promotor e outros elementos regulatórios a montante. Muito raramente uma cópia gênica processada pode reter alguma função (um *retrogene*; ver Tabela 9.7). Entretanto, como não possuem sequências importantes necessárias para expressão, a maioria das cópias gênicas processadas degenera em *pseudogenes processados* (às vezes chamados de *pseudogenes retrotranspostos*; **Figura 1B**).

Prevalência e funcionalidade dos pseudogenes

Genomas eucarióticos geralmente apresentam muitos pseudogenes. Uma explicação de longa data para sua abundância é de que a duplicação gênica seria evolutivamente vantajosa. Novas variantes de genes funcionais podem ser criadas por duplicação gênica, e os pseudogenes eram vistos como subprodutos malsucedidos dos mecanismos de duplicação. Embora alguns genomas procarióticos aparentem ter muitos pseudogenes, estes são raros em procariotos porque seus genomas são muitas vezes projetados para serem compactos.

A grande maioria do que é convencionalmente reconhecido como pseudogenes humanos são cópias de genes que codificam proteínas simplesmente porque é relativamente fácil de identificá-los (buscando mutações de mudança de quadro de leitura, de sítio de *splice*, e assim por diante). Existem mais de 8 mil cópias de pseudogenes processados de genes que codificam proteínas no genoma humano, além de mais de 4 mil pseudogenes não processados (ver o banco de dados de pseudogenes em http://www.pseudogene.org). Apenas cerca de 10% dos 21 mil genes humanos que codificam proteínas possuem pelo menos um pseudogene processado, porém genes altamente expressos tendem a ter múltiplos pseudogenes processados. Por exemplo, a proteína ribossomal citoplasmática contém 95 genes funcionais codificando 79 proteínas diferentes (16 genes são duplicados) e 2.090 pseudogenes processados.

Pseudogenes de RNA são geralmente difíceis de identificar como pseudogenes (não há fase de leitura para inspecionar, e os genes de RNA são geralmente desprovidos de íntrons). Apesar disso, cópias de pseudogenes de vários pequenos RNAs são comuns (ver tabela abaixo), notadamente se estes são transcritos pela RNA-polimerase III (genes transcritos pela RNA-polimerase III geralmente têm promotores internos).

Família de RNA	Número de genes humanos	Número de pseudogenes relacionados
snRNA U6	49	~800
snRNA U7	1	85
RNA Y	4	~1.000

Conforme será descrito na Seção 12.4, a repetição Alu, a sequência mais abundante no genoma humano, parece ter se originado pela cópia de transcritos do RNA 7SL, e muitas outras famílias de DNA altamente repetitivo intercalante em mamíferos são cópias de tRNA. Portanto, de certa maneira, pseudogenes de RNA tornaram-se as sequências mais comuns dos genomas de mamíferos.

Todos os pseudogenes estão localizados no genoma nuclear, mas eles incluem cópias defeituosas de genes que residem no genoma mitocondrial (*pseudogenes mitocondriais*). O genoma mitocondrial se originou a partir de um genoma bacteriano muito maior, e, ao longo do tempo evolutivo, grande parte do DNA do genoma mitocondrial precursor migrou em séries independentes de eventos de integração para o que hoje é conhecido como genoma nuclear. Pseudogenes do mtDNA hoje representam pelo menos 0,016% do DNA nuclear (ou cerca de 30 vezes o conteúdo do genoma mitocondrial).

A funcionalidade dos pseudogenes tem sido um debate constante, e diferentes classes de pseudogenes foram sugeridas. Um número significativo de pseudogenes (a maioria deles pseudogenes processados) é transcrito, e transcritos antissenso de pseudogenes podem regular seus genes parentais. Os pseudogenes também foram implicados diretamente na produção de siRNAs endógenos que regulam transposons, conforme descrito na Seção 9.3. Finalmente, algumas sequências de pseudogenes podem ser cooptadas para uma função diferente. Estas foram descritas como *pseudogenes exaptados*. Um exemplo disso é fornecido pelo gene *XIST*. Ele codifica um RNA não codificante que regula a inativação do cromossomo X, e dois de seus seis éxons teriam se originado a partir de uma cópia de pseudogene de um gene que codifica proteína.

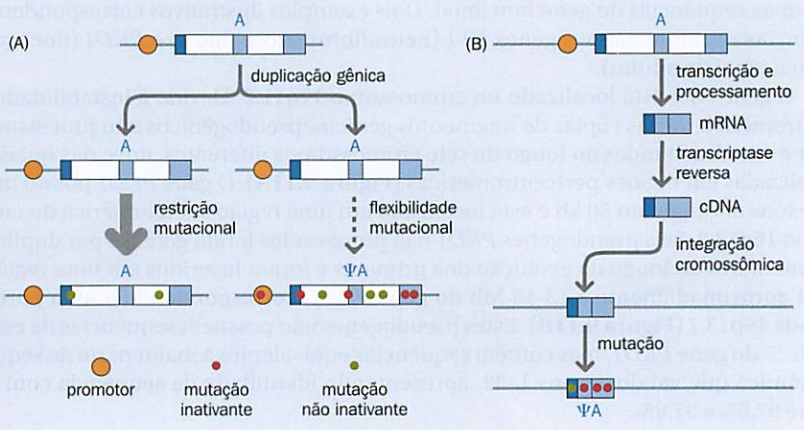


Figura 1 Origens de pseudogenes processados e não processados. (A) A cópia da sequência de DNA genômico que contém o gene A pode produzir cópias duplicadas do gene A. Uma forte pressão de seleção é aplicada para manter uma das cópias gênicas funcional (seta em negrito), mas a outra cópia fica livre para sofrer mutações (seta tracejada). Se esta cópia acumular mutações deletérias ou de inativação (círculos vermelhos), um pseudogene não processado (ψA) poderá surgir. (B) Um pseudogene processado surge depois que transcriptases reversas celulares convertem o transcrito de um gene em cDNA, o qual é então inserido de volta no genoma (ver Figura 9.12 para mais detalhes). A ausência de sequências importantes, como um promotor, geralmente resulta em uma cópia gênica inativa.

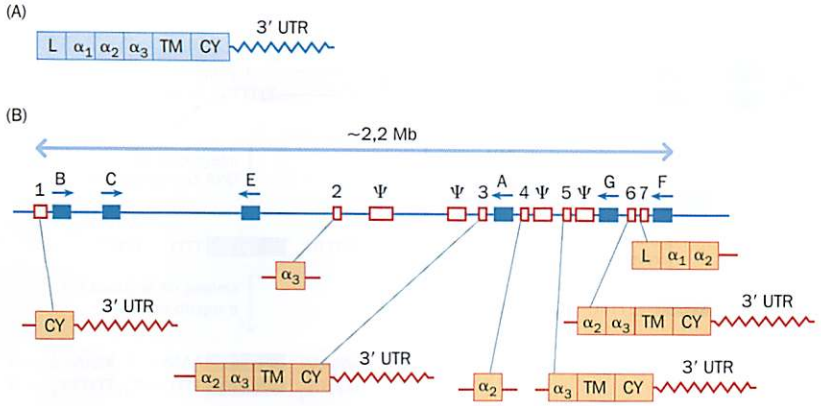


Figura 9.10 Família gênica HLA de classe I: uma família de genes agrupados com pseudogenes não processados e fragmentos gênicos. (A) Estrutura de uma molécula de mRNA de cadeia pesada de HLA de classe I. O mRNA completo possui uma sequência codificadora de polipeptídeo com uma sequência líder (L), três domínios extracelulares (α_1, α_2 e α_3), uma sequência transmembrana (TM), uma cauda citoplasmática (CY) e uma região 3′ não traduzida (3′ UTR). Os três domínios extracelulares são codificados essencialmente, cada um deles, por um éxon único. A diminuta região 5′ UTR não está representada. (B) O agrupamento gênico da cadeia pesada de HLA de classe I está localizado em 6p21.3 e compreende cerca de 20 genes. Ele inclui seis genes expressos (quadrados azuis preenchidos), quatro pseudogenes não processados completos (quadrados vermelhos grandes, identificados por ψ) e uma ampla variedade de cópias gênicas parciais (quadrados vermelhos pequenos, marcados de 1 a 7). Alguns destes são truncados na extremidade 5′ (p. ex., 1, 3, 5 e 6), outros são truncados na extremidade 3′ (p. ex., 7), e outros ainda contêm éxons únicos (p. ex., 2 e 4).

Pseudogenes processados são cópias defeituosas de um gene que apresentam apenas sequências exônicas e não possuem íntrons ou sequências promotoras a montante. Eles surgem por *retrotransposição*: transcriptases reversas celulares podem usar transcritos gênicos processados, como mRNAs, para produzir cDNA que poderá então ser integrado ao DNA cromossômico (**Figura 9.12**). Pseudogenes processados são comuns em famílias gênicas interespaçadas (ver Tabela 9.5).

Pseudogenes processados não possuem uma sequência promotora e, portanto, não são expressos. Às vezes, no entanto, a cópia de cDNA se integra em um sítio do DNA cromossômico que, por acaso, é adjacente a um promotor que pode direcionar a expressão da cópia processada do gene. Uma pressão seletiva poderá assegurar que a cópia processada do gene continue a produzir um produto gênico funcional, e neste caso ele será descrito como um **retrogene**. Uma série de retrogenes sem íntrons que apresenta padrão de expressão testículo-específico é conhecida, e estes retrogenes são geralmente homólogos autossômicos de um gene ligado ao X que contém íntrons (**Tabela 9.8**).

Uma explicação para os retrogenes poderia ser a necessidade crítica de superar a falta de expressão de certas sequências ligadas ao X nos testículos durante a meiose masculina. Durante a meiose masculina, os cromossomos X e Y pareados são convertidos em heterocromatina, formando o **corpúsculo XY**, altamente condensado e transcricionalmente inativo. Retrogenes autossômicos podem fornecer a possibilidade de síntese continuada de determinados produtos cruciais nas células dos testículos, que não são mais sintetizados pelos genes contidos no corpúsculo XY, altamente condensado.

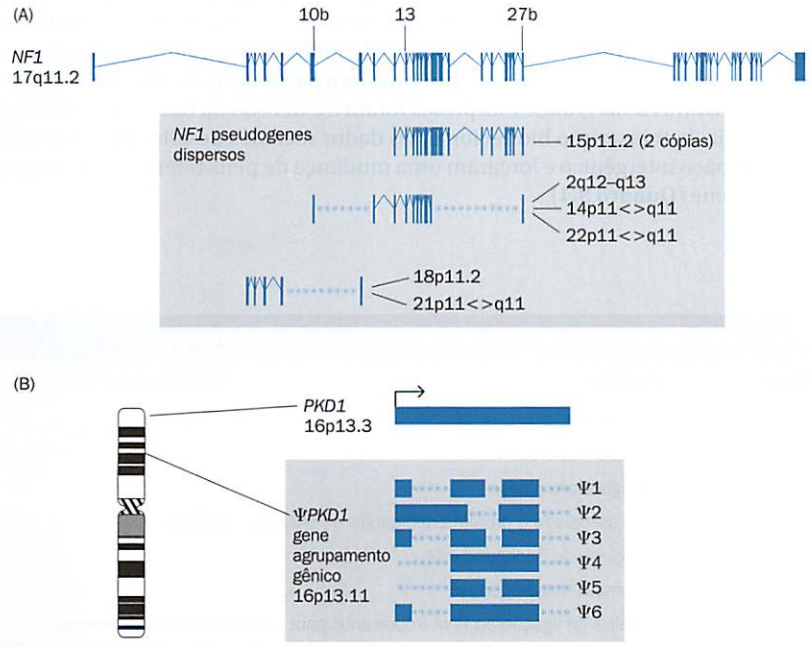


Figura 9.11 Dispersão de pseudogenes não processados de *NF1* e *PKD1* em decorrência da instabilidade das regiões pericentromérica ou subtelomérica. (A) O gene da neurofibromatose tipo 1, *NF1*, está localizado próximo ao centrômero do cromossomo humano 17. Ele ocupa uma região de 283 kb e possui 58 éxons. Os éxons são representados aqui por caixas verticais finas; os íntrons, por bifurcações conectoras (^). Cópias defeituosas altamente homólogas do gene *NF1* são encontradas em nove ou mais localizações cromossômicas diferentes, a maioria delas em regiões pericentroméricas. Cada cópia possui uma porção do gene completo, com ambos éxons e íntrons. Sete exemplos são representados aqui, como duas cópias em 15p que apresentam sequências genômicas intactas abrangendo os éxons 13 e 27b. Às vezes, rearranjos levaram à deleção de éxons e íntrons (representados por asteriscos). (B) O gene da doença policística renal *PKD1* está localizado próximo ao telômero em 16p e possui mais de 40 éxons. Em decorrência de eventos de duplicação segmentar durante a evolução dos primatas, grandes componentes deste gene foram duplicados, e seis pseudogenes de *PKD1* estão localizados em 16p13.11, com grandes blocos de sequências (representados por caixas azuis) copiados do gene *PKD1* (asteriscos representam a ausência de sequências correspondentes em *PKD1*).

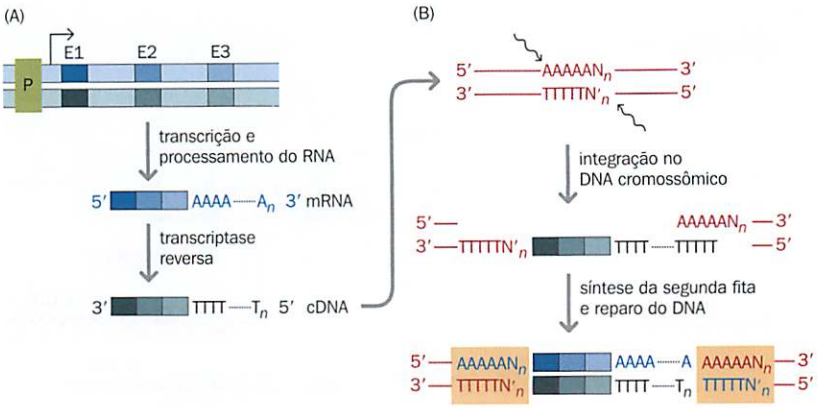


Figura 9.12 Pseudogenes processados e retrogenes originados por transcrição reversa de transcritos de RNA. (A) Neste exemplo, um gene codificador de proteína contendo três éxons (E1-E3) é transcrito a partir de um promotor a montante (P), e íntrons são removidos do transcrito para gerar um mRNA. O mRNA pode então ser naturalmente convertido em uma molécula de cDNA antissenso de fita simples por uma transcriptase reversa celular (fornecida por repetições LINE-1). (B) A integração do cDNA se dá em quebras escalonadas (indicadas por setas curvas) de sequências ricas em A, a qual pode ser auxiliada pela endonuclease de LINE-1. Se a sequência rica em A está incluída em uma extremidade 5′ disponível, esta poderá formar um híbrido com a extremidade distal da cauda poli(T) do cDNA, facilitando a síntese da segunda fita. Devido às quebras escalonadas durante a integração, a sequência inserida será flanqueada por repetições diretas curtas (sequências em caixas).

9.3 GENES DE RNA

Grande parte da atenção dirigida aos genes humanos focou nos genes que codificam proteínas porque estes sempre foram considerados a parte funcionalmente mais importante do genoma. Comparativamente, genes cujo produto final correspondem a moléculas funcionais de **RNA não codificante** (**ncRNA**) foram tão menosprezados que uma das duas sequências-rascunho do genoma humano, publicada em 2001, não continha qualquer análise de genes humanos de RNA! O RNA era visto com uma molécula que havia sido importante no início da evolução (**Quadro 9.3**), mas pensava-se que suas funções haviam sido, em grande parte, substituídas por DNA e proteínas. Até pouco tempo atrás, acreditava-se que a grande maioria das moléculas de RNA servia apenas como molécula acessória na produção de proteínas.

Os últimos anos testemunharam uma revolução no entendimento acerca da importância do RNA, e, embora o número de genes que codificam proteínas tenha sido significativamente reduzido desde que as sequências-rascunho do genoma humano foram publicadas em 2001, o número de genes de RNA tem sido constantemente revisado e aumentado. O pequeno genoma mitocondrial sempre foi considerado excepcional porque 65% (24 de 37) de seus genes correspondem a genes de RNA. Agora se começa a entender que o RNA transcrito no núcleo não é tão somente dedicado à produção de proteínas, como se pensava antes; em vez disso, ele apresenta grande diversidade funcional.

O que mudou o pensamento humano? Em primeiro lugar, classes completamente novas e desconhecidas de ncRNA foram recentemente descobertas, incluindo várias classes prolíficas de pequenos RNAs regulatórios. Em segundo lugar, análises recentes do genoma completo, usando microarranjos e sequenciamento de transcritos de alto rendimento, mostraram que pelo menos 85% e provavelmente mais de 90% do genoma humano são transcritos. Isto é, mais de 85% das posições nucleotídicas do genoma eucromático estão representadas em transcritos primários produzidos a partir de, pelo menos, uma das duas fitas do DNA. Outras duas grandes surpresas foram a extensão da transcrição multigênica e a penetrância da transcrição bidirecional. Os dados recentes desafiaram a distinção entre genes e espaço intergênico e forçaram uma mudança de pensamento radical acerca do conceito de gene (**Quadro 9.4**).

TABELA 9.8 Exemplos de retrogenes humanos desprovidos de íntrons e seus homólogos parentais que contêm íntrons

Retrogene	Homólogo contendo íntron	Produto
GK2 em 4q13	*GK1* em Xp21	glicerol cinase
PDHA2 em 4q22	PDHA1 em Xp22	piruvato desidrogenase
PGK2 em 6p12	PGK em Xq13	fosfoglicerato cinase
TAF1L em 9p13	TAF1 em Xq13	fator associado à proteína de ligação ao TATA *box*, 250 kD
MYCL1 em 1p34	MYCL2 em Xq22	homólogo do oncogene v-Myc
GLUD1 em 10q23	GLUD2 em Xq25	glutamato desidrogenase
RBMXL em 9p13	RBMX em Xq26	proteína de ligação ao RNA importante para o desenvolvimento cerebral

QUADRO 9.3 A hipótese do mundo de RNA

As proteínas não são capazes de autorreplicação, e por isso muitos geneticistas evolutivos consideram que ácidos nucleicos *autocatalíticos* devem ter surgido antes das proteínas, com capacidade de replicar-se sem a ajuda delas. A *hipótese do mundo de RNA* foi desenvolvida a partir de ideias propostas por Alexander Rich e Carl Woese, nos anos 60. Ela propõe que o RNA teria tido um papel duplo nos primeiros estágios do surgimento da vida, atuando como material genético (com capacidade de autorreplicação) e também como molécula efetora. Ambos os papéis ainda são evidentes hoje: alguns vírus têm genoma de RNA, e moléculas de RNA não codificante podem atuar como moléculas efetoras com atividade catalítica. A RNAse P, por exemplo, é uma ribozima capaz de clivar substratos de RNA sem qualquer necessidade de proteínas, e certos tipos de íntron são autocatalíticos e capazes de excisarem a si próprios dos transcritos de RNA sem a ajuda de proteínas (ver texto). Outra observação consistente com a hipótese de que o RNA teria sido o primeiro ácido nucleico é o fato de que os desoxirribonucleotídeos são sintetizados a partir de ribonucleotídeos em rotas celulares.

Além de estocar a informação genética, imagina-se que o RNA tenha sido usado subsequentemente para sintetizar proteínas a partir de aminoácidos. Diferentes RNAs, incluindo rRNA e tRNA, são centrais no processo de síntese de polipeptídeos. Muitas proteínas ribossomais podem ser deletadas sem afetar as funções do ribossomo, e a atividade essencial peptidil transferase – a enzima que cataliza a formação de ligações peptídicas – é uma ribozima. Entretanto, o RNA possui um arcabouço um tanto rígido, não sendo muito adequado como molécula efetora. As proteínas são muito mais flexíveis e também oferecem maior variedade funcional, já que os 20 aminoácidos podem ter estruturas amplamente diferentes e oferecem maior possibilidade de combinações de sequências (um decapeptídeo oferece 20^{10} ou cerca de 10^{13} diferentes sequências possíveis de aminoácidos, enquanto um decanucleotídeo apresenta 4^{10} ou cerca de 10^6 sequências diferentes possíveis).

A substituição do RNA por DNA como molécula para estocar a informação genética forneceu vantagens significativas. O DNA é muito mais estável que o RNA e, portanto, mais adequado para esta função. Seus resíduos de açúcar são desprovidos do grupo 2'-OH presente na ribose, que tornam o RNA sujeito à clivagem hidrolítica. Maior eficiência pôde ser alcançada com a separação das funções de estoque e transmissão da informação genética (DNA) e síntese de proteínas (RNA). A única coisa necessária para isso foi o desenvolvimento de uma transcriptase reversa para que o DNA pudesse ser sintetizado a partir de desoxinucleotídeos utilizando um molde de RNA.

Sabe-se, há muitas décadas, que várias classes de ncRNA ubíquas são essenciais para o funcionamento celular. Até recentemente, no entanto, estavam acostumados a pensar em ncRNA como uma série de *acessórios* necessários para transformar genes em proteínas. Os RNAs transportadores são necessários no final da rota de tradução, servindo para decodificar os códons do mRNA e fornecer os aminoácidos na ordem necessária para sua inserção nas cadeias polipeptídicas em formação. Os RNAs ribossomais são componentes essenciais dos ribossomos, as complexas fábricas ribonucleoproteicas que sintetizam proteínas.

Outros ncRNAs ubíquos eram conhecidos por atuar também em fases mais iniciais desta rota para garantir o processamento correto de precursores de mRNA, rRNA e tRNA. Várias moléculas pequenas de RNA fazem parte das ribonucleoproteínas complexas envolvidas em diferentes reações de processamento, incluindo a retirada de íntrons, a clivagem de precursores de rRNA e tRNA e modificações de base necessárias para o amadurecimento do RNA. Estes RNAs atuam como *RNAs guias*, por meio do pareamento de bases com sequências complementares no RNA precursor.

Há também o conhecimento de que alguns ncRNAs apresentam outras funções, como os RNAs envolvidos na inativação do cromossomo X e no *imprinting* e o RNA que compõe a ribonucleoproteína telomerase, necessária para a síntese do DNA dos telômeros (ver Figura 2.13). Porém, estes RNAs eram vistos como estranhas exceções.

Na última década, entretanto, houve uma revolução na maneira como é visto o RNA. Vários milhares de ncRNAs diferentes foram recentemente identificados em células animais. Muitos deles são regulados ao longo do desenvolvimento e mostraram ter papéis cruciais em uma grande variedade de processos diferentes que ocorrem em tecidos especializados ou estágios específicos do desenvolvimento. Vários ncRNAs já foram implicados no desenvolvimento de câncer e doenças genéticas.

Agora que se sabe que o genoma humano possui cerca de 20 mil genes que codificam proteínas, quase o mesmo que o nematódeo de 1 mm *Caenorhabditis elegans*, o qual tem apenas aproximadamente mil células – a questão é saber se a regulação baseada em RNA corresponde ao mecanismo-chave para explicar nossa complexidade. Certamente, a complexidade da regulação gênica baseada em RNA é maior em organismos complexos, conforme descrito no Capítulo 10. Talvez seja a hora de ver o genoma mais como uma máquina de RNA do que apenas uma máquina de proteínas.

QUADRO 9.4 Revisando o conceito de gene na era pós-genômica

Antes das análises de genomas completos, imaginava-se que um típico gene humano que codifica proteínas era bem-definido e separado de seus vizinhos por espaços intergênicos bem identificados. O gene seria dividido em vários éxons. Dirigido a partir de um promotor a montante, um transcrito primário complementar a apenas uma das fitas de DNA sofreria *splicing*. As sequências intrônicas funcionalmente sem importância (DNA lixo) seriam descartadas, permitindo a fusão das importantes sequências exônicas para a produção de um mRNA. A expressão seria regulada por sequências regulatórias próximas localizadas perto do promotor.

Esta imagem clara e acolhedora de um gene foi fustigada por uma série de complicações. Há tempos sabe-se que alguns genes nucleares são parcialmente sobrepostos a outros ou são inteiramente inseridos dentro de genes maiores. Sabia-se que produtos diferentes eram produzidos a partir de um único gene por meio de promotores alternativos, *splicing* alternativo e edição de RNA. Ocasionalmente, sequências de diferentes genes podiam ser processadas e unidas em nível de RNA, às vezes mesmo genes que estivessem em cromossomos diferentes (*trans-splicing*). Ocasionalmente, transcritos antissenso naturais eram observados e atuavam na regulação da expressão de transcritos senso de um gene. Alguns genes apresentavam elementos regulatórios distantes centenas de quilobases, às vezes dentro de outros genes.

Apesar das complicações listadas acima, os cientistas não estavam prontos para abandonar a ideia simples de gene descrita no primeiro parágrafo, até que análises de genomas inteiros derrubaram os conceitos antigos equivocados. Os achados significativos que forçaram uma reavaliação da organização do gene foram:

- *A transcrição é comum.* Mais de 85% do genoma humano eucromático são transcritos, e a transcrição multigênica é comum, de modo que a distinção entre genes e espaços intergênicos é agora muito menos aparente (**Figura 1**). Cerca de 70% dos genes humanos são transcritos a partir de ambas as fitas.
- *O DNA codificante representa menos de um quarto da fração altamente conservada (e, presumivelmente, funcionalmente importante) do genoma.* Isso sugeria que deveria haver mais sequências de DNA não codificante funcionalmente importantes do que previamente esperado. E assim ficou provado. Um número inesperadamente alto de sequências regulatórias conservadas e vários RNAs não codificantes (ncRNAs) novos foram, e continuam sendo, identificados, geralmente no interior de conhecidos íntrons de genes que codificam proteínas.

Em consequência das intensas análises dos transcriptomas de mamíferos, vários milhares de transcritos de função desconhecida foram revelados. Em geral, ncRNAs e transcritos que codificam proteínas estão sobrepostos, gerando padrões de transcrição complicados (**Figura 2**). Em alguns casos, como no transcrito imprintado *SNURF-SNRPN* em 15q12, um único transcrito contém um RNA codificante mais um RNA não codificante que são separados pela clivagem do RNA.

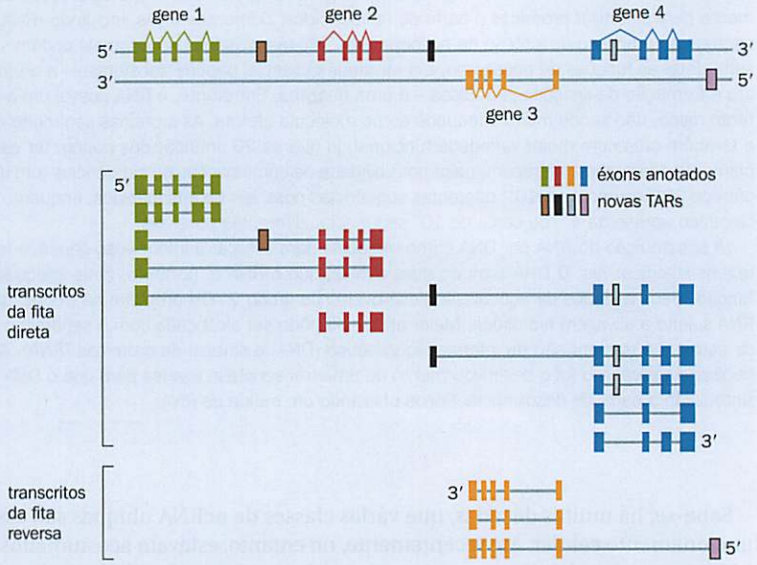


Figura 1 Indefinição dos limites do gene em nível de transcrição. No passado, esperava-se que os quatro genes no topo da figura se comportariam como unidades de transcrição individuais, discretas, não sobrepostas. Conforme demonstrado por análises recentes, a realidade é mais complicada. Uma ampla variedade de transcritos geralmente une éxons de genes vizinhos. Os transcritos frequentemente incluem sequências de regiões ativas transcricionalmente (TARs), previamente insuspeitas. [Reproduzida de Gerstein MB, Bruce C, Rozowsky JS et al. (2007) *Genome Res.* 17, 669–681. Com permissão de Cold Spring Harbor Laboratory Press.]

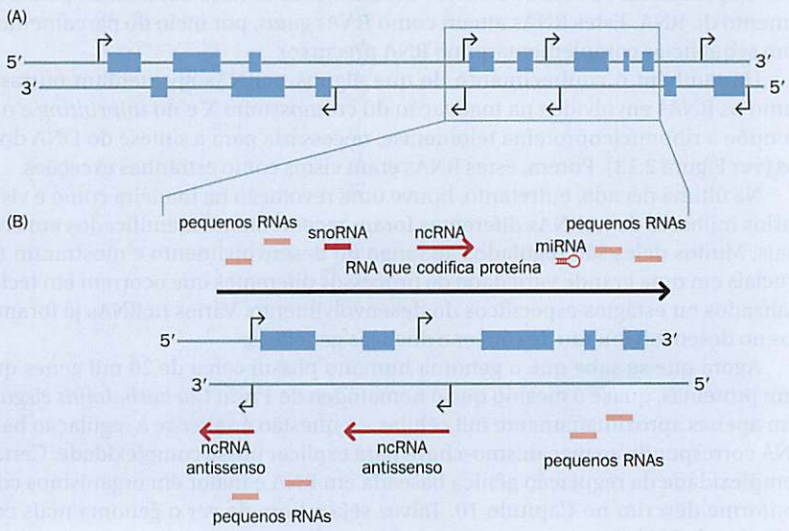


Figura 2 Complexidade transcricional extensa dos genes humanos. (A) Ambas as fitas de genes humanos são frequentemente transcritas, como representado neste agrupamento gênico hipotético. (B) Um único gene pode ter múltiplos sítios de início de transcrição (setas com ângulos para direita), bem como diversos transcritos codificantes e não codificantes intercalados. Os éxons estão representados como caixas azuis. RNAs curtos conhecidos, como os pequenos RNAs nucleolares (snoRNAs) e os microRNAs (miRNAs), podem ser processados a partir de sequências intrônicas, e novas espécies de pequenos RNAs que se agrupam próximos ao início e ao fim dos genes foram recentemente descobertas (ver o texto). [Reproduzida de Gingeras TR (2007) *Genome Res.* 17, 682-690. Com permissão de Cold Spring Harbor Laboratory Press.]

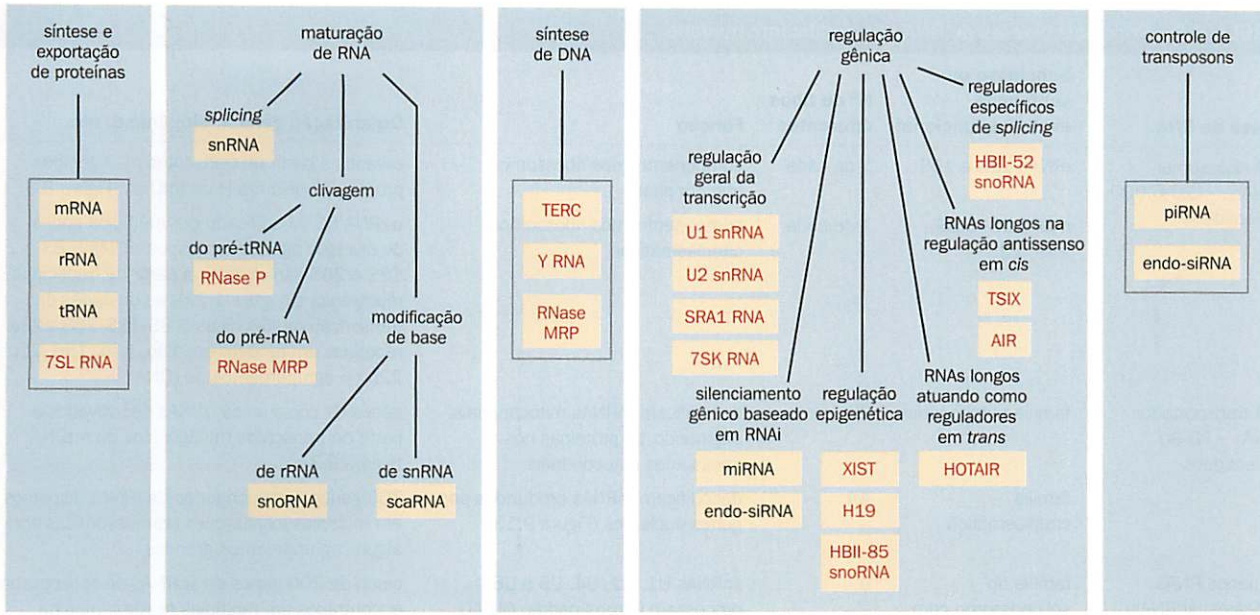


Figura 9.13 Diversidade funcional do RNA. Vários RNAs ubíquos atuam em atividades de manutenção das células e na síntese e exportação de proteínas a partir das células (utilizando o RNA 7SL, componente de RNA da partícula de reconhecimento de sinal). RNAs envolvidos na maturação de RNA incluem os pequenos RNAs nucleares (snRNAs) do spliceossomo, os pequenos RNAs nucleolares (snoRNAs), os pequenos RNAs dos corpúsculos de Cajal (scaRNAs) e duas RNA ribonucleases. A síntese do DNA dos telômeros é realizada por uma ribonucleoproteína que é formada por TERC (o componente de RNA da telomerase) e uma transcriptase reversa (ver Figura 2.13). A família de RNAs Y está envolvida na replicação do DNA cromossômico, e a RNase MRP tem papel crucial no início da replicação do mtDNA e na clivagem de precursores pré-rRNAs no nucléolo. Diferentes classes de RNA têm função regulatória na expressão gênica. Embora algumas delas tenham papéis acessórios gerais na transcrição, a expressão de RNAs regulatórios é geralmente restrita a certos tipos celulares e/ou estágios do desenvolvimento. Três classes de pequenos RNAs utilizam rotas de RNA de interferência (RNAi) para atuar como reguladores: RNAs que interagem com proteína Piwi (piRNAs) regulam a atividade de transposons em células da linhagem germinativa; microRNAs (miRNAs) regulam a expressão de genes-alvo; e RNAs de interferência curtos endógenos (endo-siRNAs) atuam como reguladores de genes e também regulam alguns tipos de transposons. Alguns membros da família snoRNA, tais como o snoRNA HBII-85, estão envolvidos na regulação gênica epigenética. Um grande número de RNAs longos está envolvido na regulação de vários genes, geralmente em nível transcricional. Alguns estão reconhecidamente envolvidos na regulação epigenética do *imprinting*, da inativação do X, e assim por diante. Famílias de RNA estão representadas aqui em preto; RNAs individuais são apresentados em vermelho.

A **Figura 9.13** oferece uma perspectiva moderna da diversidade funcional do RNA. Nesta seção foram consideradas as funções e a organização gênica das diferentes classes de RNA humano (**Tabela 9.9**). Vários bancos de dados foram recentemente desenvolvidos para documentar os dados acerca de ncRNAs (**Tabela 9.10**).

Mais de mil genes humanos codificam rRNA ou tRNA, a maioria deles localizada em grandes agrupamentos gênicos

Genes de RNA ribossomal

Além das duas moléculas de rRNA mitocondrial (rRNAs 12S e 16S), existem quatro tipos de rRNA citoplasmático, três dos quais associados à subunidade grande do ribossomo (rRNAs 28S, 5.8S e 5S) e um deles à subunidade pequena do ribossomo (rRNA 18S). Os genes que codificam o rRNA 5S ocorrem em pequenos agrupamentos gênicos, sendo o maior deles um conjunto de 16 genes no cromossomo 1q42, próximo ao telômero. Apenas poucos genes de rRNA 5S foram validados como funcionais, e há muitos pseudogenes dispersos.

Os rRNAs 28S, 5.8S e 18S são codificados por uma única unidade de transcrição multigênica (ver Figura 1.22) que é repetida em *tandem*, formando arranjos de *DNA ribossomal* de tamanho igual a megabases (cerca de 30 a 40 repetições em *tandem* ou aproximadamente 100 genes de rRNA) nos braços curtos de cada um dos cromossomos acrocêntricos humanos: 13, 14, 15, 21 e 22. Não se sabe o número exato de genes, pois os arranjos de DNA ribossomal foram excluídos do Projeto Genoma Humano devido a dificuldades técnicas em se obter um ordenamento não ambíguo dos clones de DNA sobrepostos derivados de longas regiões compostas por repetições em *tandem* muito semelhantes.

TABELA 9.9 Principais classes de RNA não codificante humano

Classe de RNA	Subclasse ou subfamília evolutiva/funcional	Nº de tipos diferentes	Função	Organização gênica, biogênese, etc.
RNA ribossomal (rRNA), ~120-5.000 nucleotídeos	rRNAs 12S e 16S	1 de cada	componentes dos ribossomos mitocondriais	clivados a partir de transcritos multigênicos produzidos pela fita H do mtDNA (Figura 9.3)
	rRNAs 5S, 5.8S, 18S, 28S	1 de cada	componentes dos ribossomos citoplasmáticos	o rRNA 5S é codificado por múltiplos genes de diversos agrupamentos; os rRNAs 5.8S, 18S e 28S são clivados a partir de transcritos multigênicos (Figura 1.22); as unidades de transcrição multigênicas 5.8S-18S-28S estão repetidas em *tandem* em 13p, 14p, 15p, 21p e 22p (= agrupamentos de rDNA)
RNA transportador (tRNA) ~70-80 nucleotídeos	família mitocondrial	22	decodificam mRNAs mitocondriais, originando 13 proteínas nos ribossomos mitocondriais	genes de cópia única; tRNAs são clivados a partir de transcritos multigênicos de mtDNA (Figura 9.3)
	família citoplasmática	49	decodificam mRNAs produzidos por genes nucleares (Figura 9.13)	700 genes e pseudogenes de tRNAs dispersos em múltiplas localizações cromossômicas com alguns agrupamentos grandes
Pequenos RNAs nucleares (snRNAs) ~60-360 nucleotídeos	família do spliceossomo com subclasses Sm e Lsm (Tabela 9.10)	9	snRNAs U1, U2, U4, U5 e U6 processam íntrons-padrão GU-AG (Figura 1.19); snRNAs U4atac, U6atac, U11 e U12 processam íntrons raros AU-AC	cerca de 200 genes de snRNAs de spliceossomo encontrados em múltiplas regiões, mas há grandes agrupamentos de genes snRNA de U1 e U2; a maioria é transcrita pela RNA pol II
	snRNAs não spliceossômicos	muitos	snRNA U7: processamento 3′ do mRNA de histonas; RNA 7SK: regulador transcricional geral; família de RNA Y: envolvida na replicação do DNA cromossômico e reguladora da proliferação celular	a maioria corresponde a genes funcionais de cópia única
Pequenos RNAs nucleolares (snoRNAs) ~60-300 nucleotídeos	classe *box* C/D (Figura 9.15A)	246	maturação de rRNA, maioria metilações de nucleotídeos 2′-*O*-ribose sítio-específicas	geralmente em íntrons de genes que codificam proteínas; múltiplas localizações cromossômicas, mas alguns genes são encontrados em múltiplas cópias em agrupamentos gênicos (tais como os agrupamentos HBII-52 e HBII-85 – Figura 11.22)
	classe H/ACA (Figura 9.15B)	94	maturação do rRNA pela modificação de uridinas,em posições específicas, em pseudoridinas	
RNA de corpúsculos de Cajal (scaRNA)		25	maturação de certas classes de snRNAs nos corpúsculos de Cajal (corpúsculos torcidos) no núcleo	geralmente no interior de íntrons de genes que codificam proteínas
RNA Ribonucleases, ~260-320 nucleotídeos		2	RNase P cliva pré-tRNAs no núcleo e na mitocôndria; RNase MRP cliva rRNAs no nucléolo e está envolvida no início da replicação do mtDNA	genes cópia única
Miscelânea de pequenos RNAs citoplasmáticos, ~80-500 nucleotídeos	BC200	1	RNA neuronal que regula a biossíntese de proteína dendrítica; originado de repetições Alu	1 gene, *BCYRN1*, em 2p16
	RNA 7SL	3	componente da partícula de reconhecimento de sinal (SRP) que medeia a inserção de proteínas secretoras no lúmen do retículo endoplasmático	três genes intimamente relacionados, agrupados em 14q22
	TERC (componente de RNA da telomerase)	1	componente da telomerase, a ribonucleoproteína que sintetiza o DNA telomérico, usando TERC como molde (Figura 2.13)	gene de cópia única em 3q26
	RNA *vault*	3	componentes das RNPs *vault* citoplasmáticas que parecem estar envolvidas na resistência a drogas	*VAULTRC1*, *VAULTRC2* e *VAULTRC3* estão agrupados em 5q31 e compartilham ~84% de identidade de sequência
	RNA Y	4	componentes da ribonucleoproteína Ro de 60 kD, um alvo importante das respostas humorais autoimunes	*RNY1*, *RNY3*, *RNY4* e *RNY5* estão agrupados em 7q36

TABELA 9.9 Principais classes de RNA não codificante humano (*continuação*)

Classe de RNA	Subclasse ou subfamília evolutiva/funcional	Nº de tipos diferentes	Função	Organização gênica, biogênese, etc.
MicroRNAs (miRNAs) ~22 nucleotídeos	> 70 famílias de miRNAs relacionados	~1.000	vários papéis importantes na regulação gênica, notadamente no desenvolvimento, e implicados em alguns tipos de câncer	ver na Figura 9.17 exemplos da organização no genoma, e na Figura 9.16, como são sintetizados
RNAs que se ligam a piwi (piRNAs) ~24-31 nucleotídeos	89 agrupamentos individuais	> 15.000	geralmente derivados de repetições; expressos apenas em células da linhagem germinativa, onde limitam atividade excessiva de transposons	89 grandes agrupamentos ao longo do genoma; agrupamentos individuais com 10–75 kb de extensão, com média de 170 piRNAs por agrupamento
Pequenos RNAs endógenos de interferência (endo-siRNAs) ~21-22 nucleotídeos	diversas	provavelmente, mais de 10.000[a]	geralmente derivados de pseudogenes, repetições invertidas, etc.; envolvidos na regulação gênica em células somáticas, e podem estar envolvidos também na regulação de alguns tipos de transposons	agrupamentos em diversas regiões do genoma
Longos RNAs não codificantes regulatórios, geralmente > 1 kb	diversas	> 3.000	envolvidos na regulação da expressão gênica; alguns estão envolvidos na expressão monoalélica (inativação do X, *imprinting*) e/ou como reguladores antissenso (Tabela 9.11)	geralmente, genes individuais de cópia única; os transcritos geralmente sofrem capeamento, *splicing* e poliadenilação, mas RNAs regulatórios antissenso são geralmente transcritos longos que não sofrem *splicing*

[a]Com base na extrapolação de dados obtidos em estudos com células de camundongos.

Genes de RNA transportador

Os 22 tRNAs mitocondriais diferentes são codificados por 22 genes tRNA no mtDNA. O Banco de Dados de tRNA Genômico (*Genomic tRNA Database*) lista mais de 500 genes de tRNA humanos que produzem tRNAs citoplasmáticos com uma especificidade do anticódon definida. Os genes podem ser classificados em 49 famílias com base na especificidade de seus anticódons (**Quadro 9.5**). Existe uma pequena correlação entre o número de genes humanos para tRNA e a frequência de aminoácidos. Por exemplo, 30 genes de tRNA especificam citeína, um aminoácido relativamente raro (o qual corresponde a cerca de 2,25% de todos os aminoácidos das proteínas humanas), mas apenas 21 genes de tRNAs especificam a prolina, que é mais abundante (com frequência de 6,10%).

Embora os genes de tRNA pareçam estar dispersos ao longo do genoma humano, mais da metade deles (273 de um total de 516) localiza-se ou no cromossomo 6 (onde muitos deles estão agrupados em uma região de 4 Mb em 6p2) ou no cromossomo 1. Além disso, 18 dos 30 tRNAs para Cys são encontrados em uma faixa de 0,5 Mb no cromossomo 7.

TABELA 9.10 Principais bancos de dados sobre RNAs não codificantes

Banco de dados	Descrição	URL
NONCODE	banco de dados integrado que inclui todos os ncRNAs, exceto rRNA e tRNA	http://www.noncode.org
Banco de dados de RNAs não codificantes	sequências e funções de transcritos não codificantes	http://biobases.ibch.poznan.pl/ncRNA
RNAdb	banco de dados abrangente de RNAs não codificantes de mamíferos	http://www.research.imb.uq.edu.au/rnadb
Rfam	famílias de RNAs não codificantes e alinhamento de sequências	http://rfam.sanger.ac.uk/
antiCODE	banco de dados de transcritos antissenso naturais	http://www.anticode.org
sno/scaRNAbase	pequenos RNAs nucleolares e pequenos RNAs específicos de corpúsculo de Cajal	http://gene.fudan.sh.cn/snoRNAbase.nsf
snoRNA-LBME-db	snoRNAs humanos	http://www-snorna.biotoul.fr/
Banco de dados de tRNA genômico	sequências de tRNA	http://lowelab.ucsc.edu/GtRNAdb/
Compilação de sequências de tRNA e de sequências de genes de tRNA	exatamente o que seu nome sugere	http://www.tRNA.uni-bayreuth.de
miRBase	sequências de miRNAs e genes-alvo	http://microrna.sanger.ac.uk/
piRNAbank	sequências deduzidas empiricamente e outras informações relativas a piRNAs de vários organismos, incluindo humanos, camundongo, rato e *Drosophila*	http://pirnabank.ibab.ac.in/

QUADRO 9.5 Especificidade dos anticódons de tRNAs citoplasmáticos eucarióticos

Não há uma correspondência de 1 para 1 entre os códons do mRNA citoplasmático e os anticódons dos tRNAs que os reconhecem. As 64 possibilidades de códons são apresentadas na **Figura 1**, juntamente com os anticódons (não modificados). Linhas horizontais unem os pares de códon-anticódon. Códons alternativos que diferem na terceira base, tendo C ou U nesta posição, podem ser reconhecidos por um único anticódon (*oscilação da terceira base*, identificada por ^). Existem três regras para decodificar códons de mRNAs citoplasmáticos:

- *Códons em grupos de dois códons.* Os códons que terminam em U/C e codificam um aminoácido diferente daqueles terminando em A/G são conhecidos como grupos de dois códons. Aqui, a oscilação da posição U/C é decodificada por um G na posição da base 5′ do anticódon do tRNA. Por exemplo, no topo à esquerda, para Phe, não há tRNA com um anticódon **A**AA para parear com o códon UU**U**, mas o anticódon **G**AA pode reconhecer ambos os códons UU**U** e UU**C** no mRNA (ver Figura 1).
- *Códons não glicina em grupos de quatro códons.* Grupos de quatro códons são aqueles nos quais U, C, A e G na terceira posição, posição oscilante, codificam o mesmo aminoácido. Neste caso, a oscilação U/C é decodificada por uma adenosina quimicamente modificada, conhecida como inosina, na posição 5′ do anticódon (caixas azuis preenchidas; ver Figura 11.31 para a estrutura da inosina). A inosina pode parear com A, C ou U. Por exemplo, na base da figura à esquerda, os códons GU**U** e GU**C** do grupo de quatro códons para valina são decodificados por um tRNA com o anticódon **A**AC, o qual é modificado para **I**AC. O anticódon **I**AC pode reconhecer cada um dos códons GU**U**, GU**C** e GU**A**. Para evitar possíveis erros de tradução, os tRNAs com inosina na posição 5′ do anticódon não podem ser usados em grupos de dois códons.
- *Códons de glicina.* O grupo de quatro códons para glicina fornece uma exceção para a regra descrita anteriormente: os códons GGU e GGC são decodificados por um anticódon **G**CC, em vez do esperado anticódon **I**CC.

Desta forma, são necessários apenas 16 anticódons para decodificar os 32 códons que terminam em U/C. O conjunto mínimo de anticódons é, portanto, 45 (64 menos 3 códons de parada, menos 16). A partir deste cálculo, poderia-se prever um total de 45 classes diferentes de tRNAs humanos. Entretanto, apesar da generalização da oscilação da terceira base, três pares de códons que terminam em U/C são reconhecidos por dois anticódons cada (ver Figura 1), e portanto há três classes extras de tRNA. Além disso, um tRNA especializado carrega um anticódon para o códon UGA (o qual normalmente funciona como códon de parada). Sob altas concentrações de selênio, este tRNA irá ocasionalmente decodificar UGA em selenocisteína, o 21º aminoácido, em um pequeno grupo de selenoproteínas. Assim, existem 45 + 3 + 1 = 49 classes diferentes de tRNAs humanos, codificados por várias centenas de genes (ver Figura 1).

Phe	UUU – AAA	-	Ser	UCU ^ AGA	11	Tyr	UAU – AUA	1
	UUC – GAA	12		UCC – GGA	-		UAC – GUA	14
Leu	UUA – UAA	7		UCA – UGA	5	stop	UAA – UUA	-
	UUG – CAA	7		UCG – CGA	4	stop	UAG – CUA	-
Cys	UGU – ACA	-	Leu	CUU ^ AAG	12	Pro	CCU ^ AGG	10
	UGC – GCA	30		CUC – GAG	-		CCC – GGG	-
stop	UGA – UCA	-(3)*		CUA – UAG	3		CCA – UGG	7
Trp	UGG – CCA	9		CUG – CAG	10		CCG – CGG	4
His	CAU – AUG	-	Arg	CGU ^ ACG	7	Ile	AUU ^ AAU	14
	CAC – GUG	11		CGC – GCG	-		AUC – GAU	3
Gln	CAA – UUG	11		CGA – UCG	6		AUA – UAU	5
	CAG – CUG	20		CGG – CCG	4	Met	AUG – CAU	20
Thr	ACU ^ AGU	10	Asn	AAU – AUU	2	Ser	AGU – ACU	-
	ACC – GGU	-		AAC – GUU	32		AGC – GCU	8
	ACA – UGU	6	Lys	AAA – UUU	16	Arg	AGA – UCU	6
	ACG – CGU	6		AAG – CUU	17		AGG – CCU	5
Val	GUU ^ AAC	11	Ala	GCU ^ AGC	29	Asp	GAU – AUC	-
	GUC – GAC	-		GCC – GGC	-		GAC – GUC	19
	GUA – UAC	5		GCA – UGC	9	Glu	GAA – UUC	13
	GUG – CAC	16		GCG – CGC	5		GAG – CUC	13
Gly	GGU – ACC	-						
	GGC – GCC	15						
	GGA – UCC	9						
	GGG – CCC	7						

Figura 1 Mais de 500 tRNAs citoplasmáticos humanos decodificam os 61 códons que especificam os 20 aminoácidos padrão. As relações entre os 64 códons possíveis (posicionados ao lado dos aminoácidos à esquerda das quatro colunas principais) e seus anticódons correspondentes (à direita das quatro colunas) estão representadas aqui. O número ao lado de cada anticódon corresponde ao número de diferentes tRNAs humanos documentados no Genomic tRNA Database (ver Tabela 9.9) que carregam aquele anticódon. Observe que 12 dos 61 anticódons que poderiam reconhecer os códons que especificam os 20 aminoácidos padrão não estão representados nos tRNAs (mostrados como traços). Isso acontece por causa da oscilação na posição da terceira base da maioria dos códons onde a terceira base é U ou C (exceto para os pares de códons AU**U**/AU**C**, AA**U**/AA**C** e UA**U**/UA**C**). O (3) marcado com um asterisco significa que há três diferentes tRNAs para selenocisteína, com um anticódon que pode reconhecer o códon UGA, o qual normalmente funciona como códon de parada. As adeninas marcadas são provavelmente uma forma modificada de adenina conhecida como inosina, na qual o grupo amina ligado ao carbono 6 é substituído por um grupo carbonil C=O.

Famílias gênicas dispersas produzem vários pequenos RNAs nucleares que facilitam a expressão gênica geral

É de conhecimento geral que várias famílias de pequenas moléculas de RNA (com 60–360 nucleotídeos de extensão), no núcleo, auxiliam a expressão gênica em geral, a maioria destas moléculas em nível de processamento pós-transcricional. Inicialmente, estes RNAs eram simplesmente denominados de *pequenos RNAs nucleares (snRNAs)* para distingui-los

TABELA 9.11 As classes de snRNAs de spliceossomo Sm E Lsm

	Classe Sm[a]	Classe Lsm
Componente do spliceossomo principal	snRNAs U1, U2, U4 e U5	snRNAs U6
Componente do spliceossomo secundário	snRNAs U11, U12, U4atac e U5	snRNAs U6atac
Estrutura	ver Figura 9.14A	ver Figura 9.14B
Transcrito por	RNA-polimerase II	RNA-polimerase III
Proteínas centrais ligadas	proteínas Sm (SmB, SmD1, SmD2, SmD3, SmE, SmF, SmG)	sete proteínas Lsm (LSM2-LSM8)
Localização	sintetizados no núcleo e exportados para o citoplasma, onde cada um se associa a sete proteínas Sm e sofre processamento nas extremidades 5′ e 3′; são, então, reimportados para o núcleo para serem novamente processados por RNA nos cospúsculos de Cajal, antes de se organizarem em grupos que realizarão funções de spliceossomo	nunca saem do núcleo; sofrem maturação no nucléolo e então se organizam em grupos que irão realizar funções de spliceossomo
Modificações de nucleotídeo sítio-específicas	realizadas por scaRNAs nos cospúsculos de Cajal do núcleo	realizadas por snoRNAs no nucléolo

[a]Embora não seja um snRNA spliceossômico, o snRNA U7 possui estrutura semelhante à classe de snRNAs Sm, e cinco de suas sete proteínas centrais são idênticas às proteínas centrais ligadas aos snRNAs Sm.

de moléculas de pré-mRNA. Sabia-se que muitos deles eram ricos em uridina e foram denominados de acordo com esta característica (snRNA U2, p. ex., não foi batizado em homenagem a uma famosa banda de *rock* irlandesa, mas simplesmente indica o segundo pequeno RNA nuclear rico em uridina que foi descrito). Da mesma forma que os rRNAs, as moléculas de snRNA ligam-se a diversas proteínas e atuam como ribonucleoproteínas (snRNPs).

Subsequentemente, vários snRNAs, incluindo alguns dos primeiros a serem classificados, foram implicados no processamento pós-transcricional de precursores de rRNA no nucléolo; estes foram, portanto, reclassificados como *pequenos RNAs nucleolares* (*snoRNAs*), por exemplo, snoRNAs U3 e U8. Mais recentemente, a distribuição dos membros de cada classe foi baseada na classificação estrutural e funcional destas moléculas.

Um terceiro grupo de pequenos RNAs, semelhantes aos snoRNAs porém restritos aos corpúsculos de Cajal (também conhecidos como *corpúsculos torcidos*), estruturas nucleares discretas intimamente associadas à maturação de snRNPs, foi identificado. Estes foram designados *pequenos RNAs de corpúsculos de Cajal* (*scaRNAs*). Centenas de genes humanos, a maioria deles dispersa, produzem snRNAs e snoRNAs, e há várias centenas de pseudogenes associados.

Genes de pequenos RNAs nucleares (snRNAs) do spliceossomo

Os nove snRNAs humanos do spliceossomo variam em comprimento de 106 a 186 nucleotídeos e se ligam a um anel de sete proteínas centrais. Os snRNAs U1, U2, U4, U5 e U6 operam no spliceossomo principal para processar íntrons convencionais GU-AG (ver Figura 1.19). Os snRNAs U4atac, U6atac, U11 e U12 formam uma porção de um spliceossomo secundário que processa íntrons raros AU-AC. Cada uma das snRNPs do spliceossomo contém sete proteínas centrais que são idênticas dentro de uma subclasse e um conjunto único de proteínas específicas de snRNP. A subclasse Lsm é composta apenas pelos snRNAs U6 e U6atac; os outros snRNAs do spliceossomo pertencem à subclasse Sm (**Tabela 9.11** e **Figura 9.14**).

Mais de 70 genes especificam snRNAs usados no spliceossomo principal. Eles incluem 44 genes identificados codificando snRNA U6, e 16 especificando snRNA U1. Existe alguma evidência de agrupamento entre eles. Múltiplos genes de snRNA U2 são encontrados no cromossomo 17q21-q22, mas o número de cópias é variável; um agrupamento de cerca de 30 genes de snRNA U1 está localizado em 1p36.1.

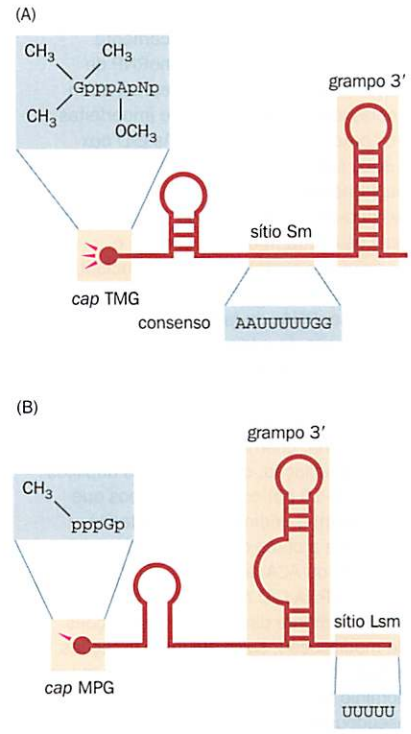


Figura 9.14 Estruturas dos snRNAs do spliceossomo tipo Sm e tipo Lsm. (A) snRNAs tipo Sm contêm três elementos de reconhecimento importantes: um *cap* de 5′-trimetilguanosina (TMG), um sítio de ligação à proteína Sm (sítio Sm) e uma estrutura haste-alça (grampo) na extremidade 3′. O sítio Sm e os elementos da haste em 3′ são necessários ao reconhecimento pelo complexo de sobrevivência de neurônios motores (SMN) para sua montagem em ribonucleoproteínas (RNPs) centrais estáveis. O sítio consenso Sm direciona a montagem de um anel de sete proteínas Sm centrais (ver Tabela 9.10). O *cap* TMG e as proteínas Sm montadas são necessárias para o reconhecimento pela maquinaria de importação nuclear. (B) snRNAs tipo Lsm contêm um *cap* de 5′-monometilfosfato guanosina (MPG) e um grampo em 3′, terminando em uma cauda de resíduos de uridina (o sítio Lsm) que é ligada por sete proteínas Lsm centrais.

Genes de pequenos RNAs nucleares não spliceossômicos

Nem todos os snRNAs do nucleoplasma atuam como parte dos spliceossomos. Tanto o snRNA U1 como o U2 também têm funções não spliceossômicas. O snRNA U1 é necessário para estimular a transcrição pela RNA-polimerase II. O snRNA U2, por sua vez, estimula o alongamento transcricional pela RNA-polimerase II. Vários outros pequenos RNAs nucleares com função não spliceossômica foram bem estudados. Eles tendem a ser genes cópia-única, mas há muitos pseudogenes associados. Três exemplos são dados a seguir:

- O U7 é um snRNA de 63 nucleotídeos dedicado ao processamento especializado da extremidade 3′ sofrido pelo mRNA de histona, o qual, excepcionalmente, não é poliadenilado.
- O RNA 7SK é um RNA de 331 nucleotídeos que atua como um regulador negativo do fator de alongamento da RNA-polimerase II P-TEFb.
- A família RNA Y consiste em três pequenos RNAs (com menos de 100 nucleotídeos) que estão envolvidos na replicação do DNA cromossômico e funcionam como reguladores da proliferação celular.

Genes de pequenos RNAs nucleolares (snoRNA)

Os snoRNAs têm tamanho entre 60 e 300 nucleotídeos e foram inicialmente identificados no nucléolo, onde guiam modificações nucleotídicas em posições específicas do rRNA. Eles fazem isso por meio da formação de pequenos complexos diméricos com uma sequência do rRNA que contém o nucleotídeo-alvo. Existem duas grandes subfamílias destes genes. Os snoRNAs H/ACA guiam pseudouridilações sítio-específicas (uridina é isomerizada em pseudouridina em 95 posições diferentes no pré-rRNA). Os snoRNAs *box* C/D guiam metilações sítio-específicas em 2′-*O*-ribose (há de 105 a 107 variações desta metilação no rRNA). Os snoRNAs únicos especificam uma, ou no máximo duas, modificações de base (**Figura 9.15**).

Pelo menos 340 genes humanos de snoRNAs foram descritos até o momento, mas pode haver muitos outros, pois snoRNAs são surpreendentemente difíceis de serem identificados com o uso de abordagens de bioinformática. A grande maioria deles está localizada no interior de íntrons de genes maiores que são transcritos pela RNA-polimerase II. Estes snoRNAs são produzidos pelo processamento do RNA intrônico, e, portanto, a regulação de sua síntese é acoplada àquela do gene hospedeiro. Muitos genes de snoRNAs são genes de cópia única dispersos. Outros, no entanto, ocorrem em agrupamentos. Por exemplo, a grande unidade de transcrição imprintada *SNURF-SNRPN* em 15q12 contém seis tipos diferentes de genes de snoRNA *box* C/D, dois dos quais estão presentes em grandes agrupamentos gênicos: um contém aproximadamente 45 genes de snoRNA HBII-52 quase idênticos, e o outro possui 29 genes de snoRNA HBII-85 (ver Figura 11.22).

A maioria dos snoRNAs é ubiquamente expressa, mas alguns são tecido-específicos. Por exemplo, os seis tipos de gene de snoRNA da unidade de transcrição *SNURF-SNRPN* são predominantemente expressos no cérebro apenas a partir do cromossomo 15 paterno. Genes de snoRNA que não possuem sequências complementares a sequências de rRNA

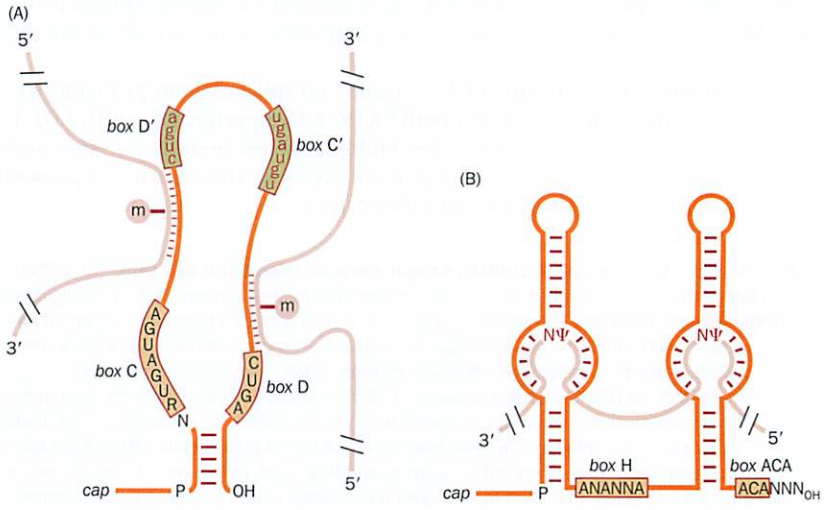


Figura 9.15 Estrutura e função dos snoRNAs. (A) Os snoRNAs da família CD *box* guiam as modificações do tipo 2′-*O*-metilação. Os motivos C e D do *box* e um pequeno grampo 5′, 3′-terminal formado por pareamento de bases entre as fitas (representado como uma série de linhas horizontais vermelhas curtas) constituem um motivo estrutural do tipo *kink-turn* que é especificamente reconhecido pela proteína snoRNP de 15,5 kD. Os *boxes* C′ e D′ representam cópias internas e geralmente imperfeitas dos *boxes* C e D. Os snoRNAs C/D *box* e seus substratos de RNA formam uma dupla-hélice de 10-21 pb na qual o resíduo-alvo a ser metilado (representado aqui pela letra m em um círculo) é posicionado exatamente cinco nucleotídeos antes do *box* D ou D′. R representa purina. (B) Os snoRNAs da família H/ACA *box* guiam a conversão de uridina em pseudouridina. Estes RNAs dobram-se em uma estrutura de grampo-dobradiça-grampo-cauda. Um ou ambos os grampos contêm uma alça interna, chamada de bolso de pseudouridilação, que forma dois dúplices curtos (3-10 pb) com nucleotídeos que flanqueiam a uridina não pareada (ψ) localizada a cerca de 15 nucleotídeos do *box* H ou ACA do snoRNA. Embora cada snoRNA C/D *box* e H/ACA *box* possa potencialmente direcionar duas reações de modificação, salvo algumas exceções, a maioria dos snoRNAs possui apenas um domínio funcional para 2′-*O*-metilação ou pseudouridilação.

provavelmente desempenham funções distintas das até o momento discutidas. Por exemplo, o snoRNA HBII-52 possui uma sequência de 18 nucleotídeos que é perfeitamente complementar a uma sequência interna do gene *HTR2C* (receptor 2c de serotonina) em Xp24 e regula o *splicing* alternativo deste gene. Os snoRNAs HBII-85 vizinhos foram recentemente implicados na patogênese da síndrome de Prader-Willi (OMIM 176270).

Genes de pequenos RNAs de corpúsculos de Cajal

Os scaRNAs são semelhantes aos snoRNAs e desempenham um papel similar na maturação do RNA, mas seus alvos são os snRNAs spliceossômicos, que promovem modificações sítio-específicas em precursores de snRNAs spliceossômicos nos corpúsculos de Cajal do núcleo. Existem pelo menos 25 genes humanos, cada um especificando um tipo de scaRNA. Assim como os genes de snoRNA, os genes de scaRNA estão localizados no interior de íntrons de genes transcritos pela RNA-polimerase II.

Aproximadamente mil microRNAs humanos diferentes regulam conjuntos complexos de genes-alvo por meio do pareamento de bases com os transcritos de RNA

Além do tRNA, conhece-se já há algum tempo uma ampla variedade de outros RNAs citoplasmáticos moderadamente pequenos (com 80 a 500 nucleotídeos). Por exemplo, a enzima telomerase, que sintetiza o DNA dos telômeros (ver Figura 2.13), tem um componente proteico, o TERT (transcriptase reversa da telomerase), e também um componente de RNA, o TERC, que é sintetizado a partir de um gene de cópia única no cromossomo 3q26.2 (ver Tabela 9.8 para outros exemplos). No início dos anos 2000, descobriu-se uma nova família de pequenos RNAs regulatórios, denominados **microRNAs** (**miRNAs**), que também operam no citoplasma.

Os microRNAs possuem tamanho de apenas 21 a 22 nucleotídeos em média e foram inicialmente negligenciados nas análises do genoma humano. Os primeiros miRNAs animais a serem descritos foram identificados em organismos modelos tais como o nematódeo (*C. elegans*) e a mosca-da-fruta (*Drosophila melanogaster*) por investigadores que estudavam fenômenos relacionados ao **RNA de interferência** (**Quadro 9.6**), uma forma natural de regulação gênica que protege as células da propagação prejudicial de vírus e transposons.

Devido à alta conservação de muitos miRNAs durante a evolução, miRNAs de vertebrados foram rapidamente identificados, e os primeiros miRNAs de humanos foram descritos no início dos anos 2000. Os miRNAs regulam a expressão de conjuntos selecionados de genes-alvo pelo pareamento de bases com seus transcritos. Geralmente, os sítios de ligação se localizam na região 3′ não traduzida das sequências-alvo do mRNA, e o miRNA ligado inibe a tradução, reduzindo desta forma a expressão do gene-alvo.

A síntese dos miRNAs envolve a clivagem de RNAs precursores por ribonucleases RNase III núcleo-específicas e citoplasma-específicas, que se ligam de maneira precisa a RNAs de dupla-fita e clivam estas moléculas. O transcrito primário, chamado *pri-miRNA*, tem repetições invertidas posicionadas em regiões próximas cujas bases pareiam para formar um grampo de RNA que é inicialmente clivado a partir do transcrito primário por uma RNase III nuclear (conhecida como Rnasen ou Drosha) para produzir um pequeno pré-miRNA de dupla-fita, o qual é transportado para fora do núcleo (**Figura 9.16**). Uma RNase III citoplasmática chamada dicer cliva o pré-miRNA para gerar um miRNA dúplex contendo dinucleotídeos 3′ livres.

Um complexo de silenciamento específico induzido por RNA (RISC) que contém a endorribonuclease argonauta se liga ao dúplex de miRNA e favorece o desenrolamento do miRNA dupla-fita. A proteína argonauta, então, degrada uma das fitas (a *fita passageira*), mantendo o miRNA maduro fita simples (conhecida como *fita guia*) ligado à argonauta. O miRNP maduro associa-se com transcritos de RNA que têm sequências complementares à fita guia. A ligação do miRNA ao transcrito-alvo normalmente envolve um número significativo de bases malpareadas. Consequentemente, um miRNA típico pode silenciar a expressão de centenas de genes-alvo da mesma forma que um fator de transcrição tecido-específico pode afetar a expressão de múltiplos genes-alvo ao mesmo tempo – ver a seção de alvos da base de dados miRBase listada na Tabela 9.9.

Para identificar outro genes miRNA, novos programas computacionais de bioinformática foram desenvolvidos com a capacidade de triar sequências genômicas. Em meados de 2009, mais de 700 genes humanos de miRNA haviam sido identificados e experimentalmente validados, e análises genômicas comparativas indicam que o número de

QUADRO 9.6 RNA de interferência como um mecanismo de defesa celular

O *RNA de interferência* (*RNAi*) é um mecanismo evolutivamente antigo utilizado em animais, plantas e até mesmo em fungos unicelulares para proteger as células contra vírus e elementos transponíveis. Tanto vírus como elementos transponíveis ativos podem produzir moléculas longas de RNA dupla-fita, pelo menos de maneira transiente durante seu ciclo de vida. Moléculas longas de RNA dupla-fita não são normalmente encontradas nas células e, em muitos organismos, elas disparam uma via de RNA de interferência. Uma endorribonuclease citoplasmática chamada dicer cliva o RNA longo em uma série de moléculas curtas de RNA dupla-fita conhecidas como **pequenos RNAs de interferência** (**siRNA**). Os siRNAs possuem em média 21 pb, mas clivagens assimétricas levam à existência de dois nucleotídeos livres "pendurados" na extremidade 3′ (**Figura 1**).

Os dúplices siRNA são ligados por diferentes complexos que contêm uma endorribonuclease tipo argonauta (Ago) e algumas outras proteínas. A seguir, as duas fitas de RNA são separadas, e uma delas é degradada pela argonauta, deixando uma molécula de RNA fita simples ligada ao complexo argonauta. O complexo argonauta é então ativado; o RNA fita-simples guiará o complexo argonauta ao seu alvo pelo pareamento de bases com sequências de RNA complementares na célula.

Um tipo de complexo argonauta é conhecido com complexo silenciador induzido por RNA (RISC). Neste caso, após a ligação do RNA guia fita simples a uma molécula de RNA complementar longa, de fita simples, a enzima argonauta irá clivar o RNA, causando sua degradação. RNAs virais e de transposons podem ser inativados desta maneira.

Outra classe de complexo argonauta é o complexo de silenciamento transcricional induzido por RNA (RITS). Aqui, o RNA guia fita simples se liga a transcritos de RNA complementares assim que eles emergem após a transcrição pela RNA-polimerase II. Isso faz o complexo RITS se posicionar em uma parte específica do genoma, atraindo proteínas tais como histona-metiltransferases (HMTs) e, às vezes, DNA-metiltransferases (DNMTs), as quais modificam histonas covalentemente nesta região. Este processo, por fim, causa a formação e a difusão de heterocromatina. Em alguns casos, o complexo RITS pode induzir a metilação do DNA. Como consequência, a expressão gênica pode ser silenciada por longos períodos, limitando, por exemplo, a atividade de transposons.

Embora as células de mamíferos tenham rotas de RNA de interferência, a presença de moléculas de RNA dupla-fita dispara uma resposta mediada por interferon que causa silenciamento gênico inespecífico e morte celular. Isso será descrito no Capítulo 12, quando for considerada a utilização de RNA de interferência como uma ferramenta experimental para produzir o silenciamento de genes-alvo pré-selecionados. Nestes casos, moléculas de RNA dupla-fita curtas sintetizadas artificialmente são usadas para provocar silenciamento gênico baseado em RNAi.

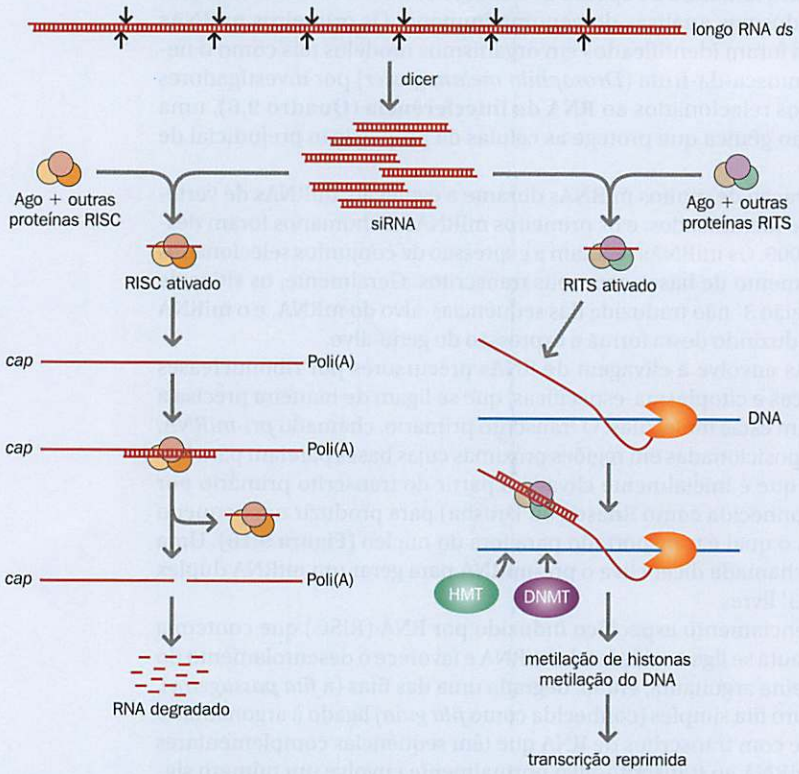


Figura 1 RNA de interferência. Moléculas longas de RNA dupla-fita (*ds*) são clivadas pela dicer citoplasmática, produzindo siRNAs. Os dúplices siRNAs são ligados por complexos argonauta, que separam o dúplex e degradam uma das fitas, produzindo um complexo ativo com uma fita simples de RNA. Por pareamento de bases com sequências complementares de RNA, o siRNA guia o complexo argonauta para reconhecer sequências-alvo. Complexos RISC ativados clivam qualquer fita de RNA que seja complementar à molécula de siRNA ligada a eles. O RNA clivado é rapidamente degradado. Complexos RITS ativados utilizam seus siRNAs para parear com qualquer RNA complementar recém-sintetizado e então atraem proteínas tais como histona-metiltransferases (HMTs) e às vezes DNA-metiltransferases (DNMTs), que podem modificar a cromatina para reprimir a transcrição.

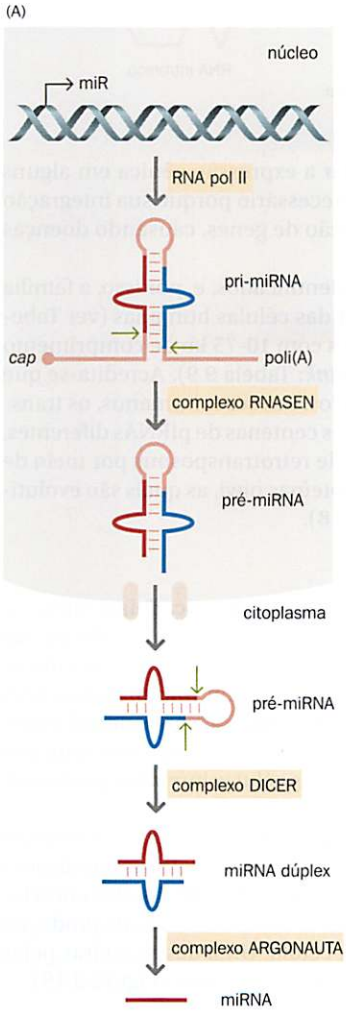
tais genes tende a aumentar. Alguns dos genes de miRNA têm promotores individuais próprios; outros fazem parte de um agrupamento de miRNA e são clivados a partir de uma unidade de transcrição comum multi-miRNA (**Figura 9.17A**). Outra classe de genes de miRNA envolve genes que formam parte de uma unidade de transcrição composta que é dedicada à produção de, além dos miRNA, outras moléculas: outro tipo de ncRNA (**Figura 9.17B**) ou uma proteína (**Figura 9.17C**).

Muitos milhares de piRNAs diferentes e siRNAs endógenos suprimem a transposição e regulam a expressão gênica

A descoberta dos miRNAs foi inesperada, mas posteriormente ficou claro que representam um pequeno componente de um conjunto enorme de diferentes pequenos RNAs regulatórios produzidos em células animais. Em mamíferos, duas classes adicionais de pequenos RNAs regulatórios foram inicialmente descritas em 2006, e estas têm sido intensamente estudadas. Como grandes números de diferentes variedades destes RNAs são gerados a partir de múltiplas regiões diferentes do genoma, o sequenciamento em larga escala foi necessário para diferenciá-los.

RNAs que interagem com a proteína Piwi

Moléculas de RNA que interagem com a proteína Piwi (piRNAs) foram encontradas em uma ampla gama de eucariotos. Elas são expressas em células da linhagem germinativa de mamíferos e possuem entre 24-31 nucleotídeos; acredita-se que tenham um papel importante limitando a transposição por retrotransposons em células da linhagem



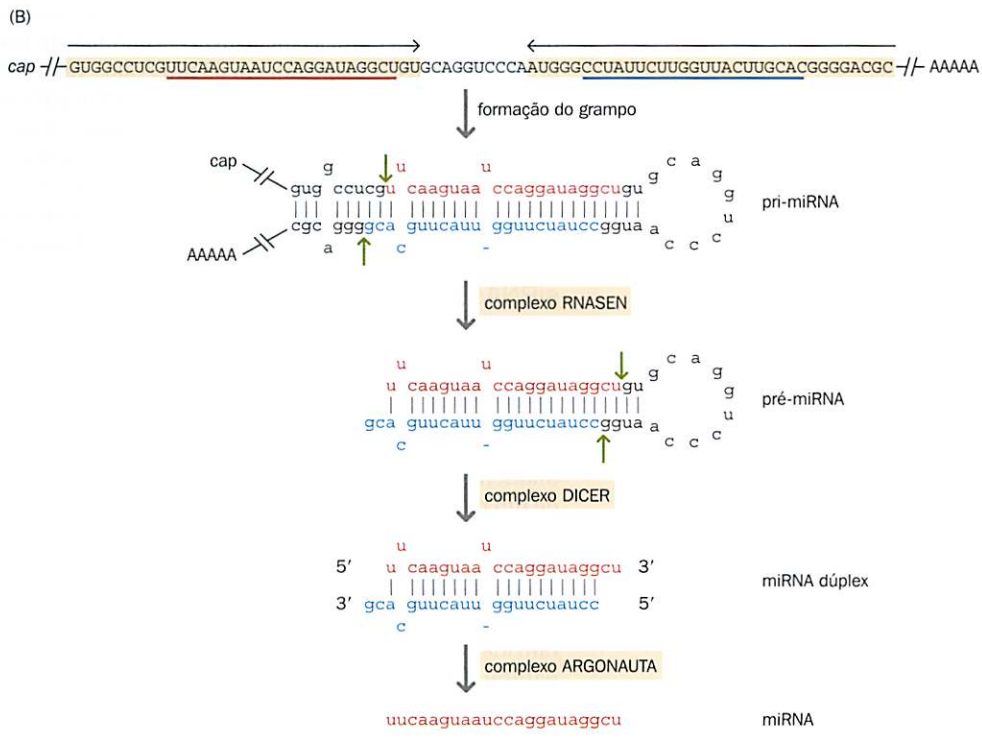


Figura 9.16 Síntese de miRNA humano. (A) Esquema geral. O transcrito primário, pri-miRNA, tem um *cap* em 5′ (m^7GpppG) e uma cauda de poli(A) em 3′. Os precursores pri-miRNA possuem uma estrutura de RNA dupla-fita proeminente (grampo de RNA), e seu processamento ocorre por meio da ação de uma série de complexos de ribonucleases. No núcleo, a Rnasen, homóloga humana da Drosha, cliva o pri-miRNA para liberar o grampo de RNA (pré-miRNA); este é então exportado para o citoplasma, onde será clivado pela enzima dicer para produzir um dúplex de miRNA. O dúplex de RNA é ligado a um complexo argonauta, e a hélice é desfeita; uma das fitas (a *passageira*) é degradada pela ribonuclease argonauta, deixando o miRNA maduro (a fita *guia*) ligado ao argonauta. miR, gene de miRNA. (B) Um exemplo específico: a síntese do miRNA humano miR-26a1. Repetições invertidas (representadas como sequências sublinhadas, cobertas por setas longas) no pri-miRNA sofrem pareamento de bases para formar um grampo, geralmente com alguns pareamentos não exatos. As sequências que darão origem à fita guia madura são representadas em vermelho; aquelas que representam a fita passageira estão em azul. A clivagem por ambas as enzimas Drosha e dicer humana (setas verdes) é geralmente assimétrica, produzindo um dúplex de RNA com dinucleotídeos 3′ livres.

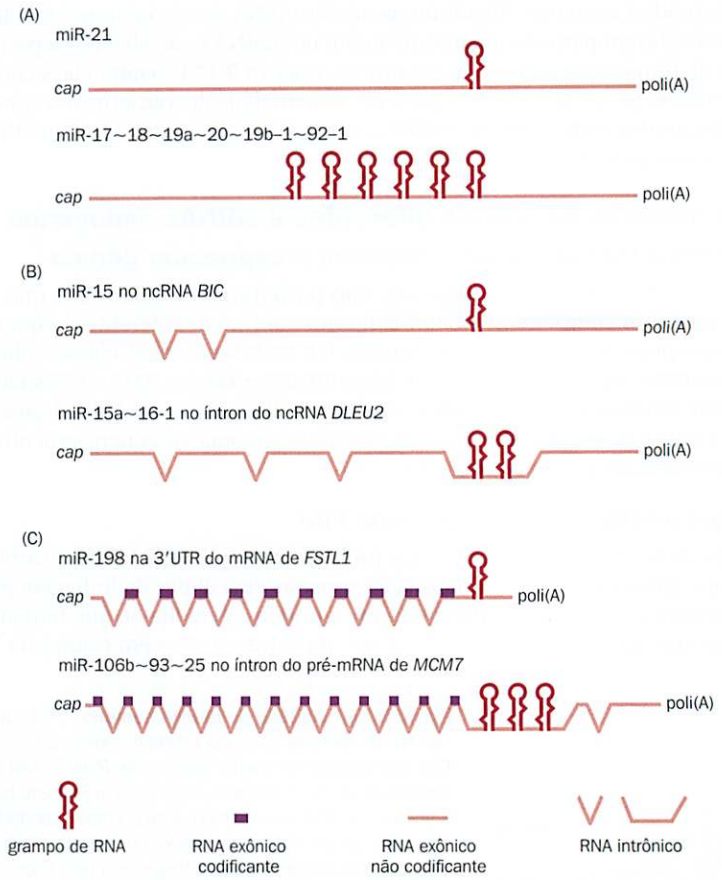


Figura 9.17 A estrutura de pri-miRNAs humanos. (A) Exemplos de transcritos que são utilizados exclusivamente para a produção de miRNAs: miR-21 é produzido a partir de um único grampo contido em um transcrito primário de RNA; um transcrito multigênico único contendo seis grampos será adequadamente clivado para produzir seis miRNAs, denominados miR-17, miR-18, miR-19a, e assim por diante. (B,C) Exemplos de miRNAs, que são cotranscritos com um gene contendo ou (B) um RNA não codificante longo (ncRNA) ou (C) um polipeptídeo. Em cada parte, o exemplo superior representa miRNAs individuais localizados em (B) um éxon de um ncRNA (miR-155) e (C) na região 3′ não traduzida (UTR) no interior de um éxon terminal de um mRNA (miR-198). Os exemplos inferiores mostram múltiplos miRNAs localizados no interior de sequências intrônicas de (B) um ncRNA (miR-15a e 16-1) e (C) um pré-miRNA (miR-106b, miR-93 e miR-25). *Cap*, $m^7G(5')ppp(5')G$. [Adaptada de Du T & Zamore PD (2005) *Development* 132, 4.645–4.652. Com permissão de Company of Biologists.]

germinativa de mamíferos, mas também podem regular a expressão gênica em alguns organismos. O controle da atividade dos transposons é necessário porque sua integração em novas regiões do genoma pode interferir com a função de genes, causando doenças genéticas e câncer.

Mais de 15 mil piRNAs humanos diferentes foram identificados, e, por isso, a família dos piRNAs está entre as mais diversas famílias de RNA das células humanas (ver Tabela 9.8). Os piRNAs mapeiam em 89 intervalos genômicos com 10-75 kb de comprimento (para mais informações, ver o banco de dados piRNA*bank*; Tabela 9.9). Acredita-se que sejam clivados a partir de grandes transcritos multigênicos. Em seres humanos, os transcritos multi-piRNA contêm sequências que atingem várias centenas de piRNAs diferentes.

Supõe-se que os piRNAs reprimam a transposição de retrotransposons por meio de uma via de RNA de interferência, por associação com proteínas piwi, as quais são evolutivamente relacionadas às proteínas argonauta (**Figura 9.18**).

siRNAs endógenos

Moléculas de RNA dupla-fita longas, em células de mamíferos, desencadeiam silenciamento gênico inespecífico por meio de rotas de interferon. No entanto, a transfecção de dúplices de siRNA sintéticos exógenos ou de grampos de RNA curtos induz o silenciamento mediado por RNAi de genes específicos que contêm elementos em comum com o RNA exógeno. Conforme será visto no Capítulo 12, esta é uma ferramenta experimental extremamente importante que pode gerar informações valiosas acerca das funções celulares de um determinado gene. Recentemente, verificou-se que as células humanas produzem *siRNAs endógenos* (*endo-siRNAs*) naturalmente.

Em mamíferos, a análise de endo-siRNAs mais abrangente foi realizada em oócitos de camundongos. Como acontece com os piRNAs, os endo-siRNAs estão entre as populações de RNA mais variadas da célula (várias dezenas de milhares de endo-siRNAs diferentes foram identificadas em oócitos de camundongos). Eles surgem como resultado da produção natural de quantidades limitadas de RNA dupla-fita na célula. Uma das maneiras pelas quais isso ocorre envolve a transcrição ocasional de alguns pseudogenes (**Figura 9.19**).

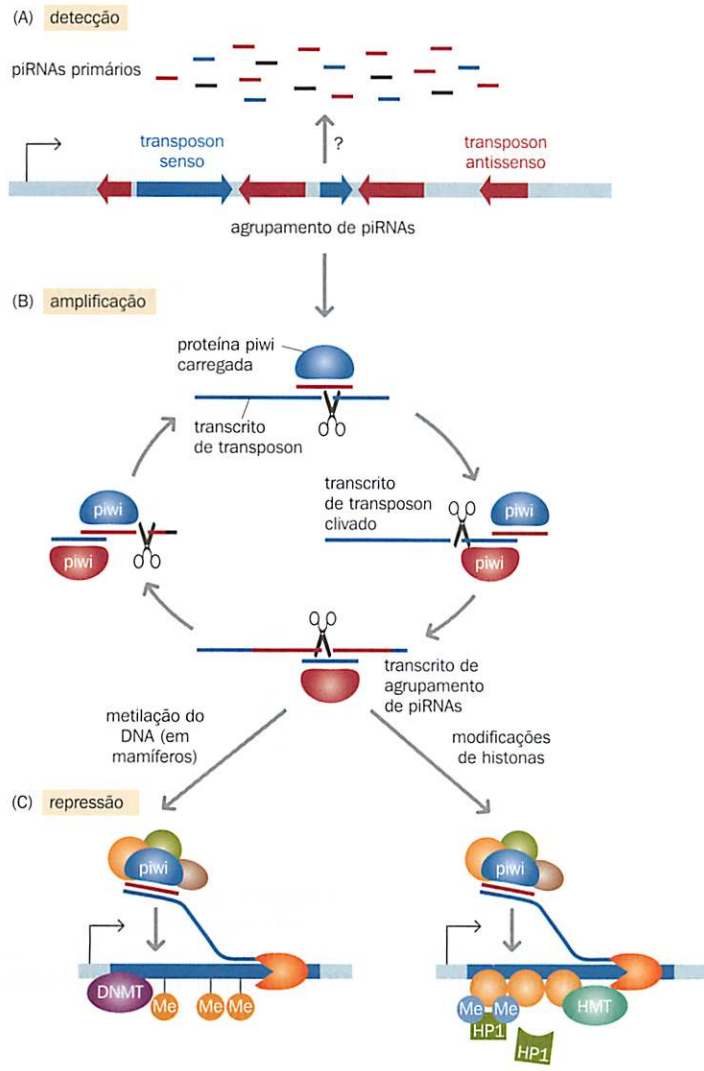


Figura 9.18 Silenciamento de transposons baseado em piRNAs em células animais. (A) piRNAs (RNAs que interagem com proteínas piwi) primários possuem 24-31 nucleotídeos e são processados a partir de precursores longos transcritos de *loci* definidos chamados agrupamentos de piRNAs. Qualquer transposon inserido em orientação inversa no agrupamento de piRNAs pode originar um piRNA antissenso (representado em vermelho). (B) piRNAs antissenso são incorporados em uma proteína piwi e dirigem sua atividade de clivagem em transcritos de transposons senso. O produto da clivagem da extremidade 3′ é ligado a outra proteína piwi e cortado até atingir o tamanho de um piRNA. Este piRNA senso é, por sua vez, usado para clivar transcritos do agrupamento de piRNAs, gerando mais piRNAs antissenso. (C) piRNAs antisssenso direcionam os complexos piwi ao cDNA para a metilação de DNA (esquerda) e/ou modificações de histonas (direita). DNMT, DNA-metiltransferase; HMT, histona-metiltransferase; HP1: proteína 1 de heterocromatina. [De Girard A & Hannon GJ (2007) *Trends Cell Biol*. 18, 136–148. Com permissão de Elsevier.]

Mais de 3 mil genes humanos sintetizam uma grande variedade de grandes e médios RNAs regulatórios

Milhares de diferentes ncRNAs longos, geralmente com muitas quilobases de extensão, parecem também ter papéis regulatórios em células animais. Estes incluem transcritos antissenso que normalmente não sofrem *splicing* e podem regular transcritos senso sobrepostos, além de uma variedade de ncRNAs longos semelhantes a mRNAs que são submetidos à modificação do quepe, *splicing* e poliadenilação, mas não parecem codificar qualquer polipeptídeo, embora alguns contenham ncRNAs internos como snoRNAs e piRNAs. As funções da grande maioria dos ncRNAs semelhantes a mRNAs são desconhecidas. Algumas destas moléculas, entretanto, são reconhecidamente tecido-específicas e estão envolvidas na regulação gênica. Recentemente, em um esforço sistemático para identificar ncRNAs longos, 3.300 ncRNAs humanos distintos foram identificados em associação com complexos modificadores da cromatina, assim afetando a expressão gênica.

Alguns ncRNAs longos semelhantes a mRNA que estão envolvidos em regulação epigenética foram extensivamente estudados. O gene *XIST* codifica um ncRNA longo que regula a inativação do cromossomo X, processo pelo qual um dos dois cromossomos X é aleatoriamente selecionado para ser condensado em fêmeas de mamíferos, com grandes regiões tornando-se transcricionalmente inativas. Muitos outros ncRNAs longos, tais como o RNA *H19*, estão implicados na repressão da transcrição do alelo paterno ou materno em diversas regiões autossômicas (*imprinting*). Estes ncRNAs semelhantes a mRNA são muitas vezes regulados por genes que produzem transcritos de ncRNAs antissenso muito longos, os quais geralmente não sofrem *splicing* (a **Tabela 9.12** mostra exemplos).

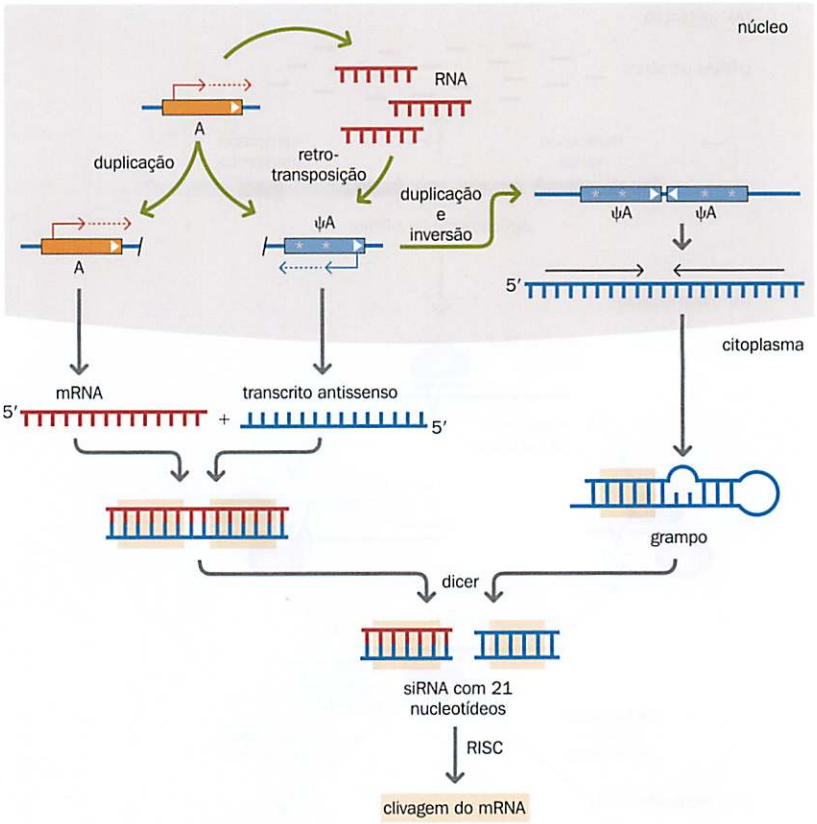


Figura 9.19 Pseudogenes podem regular a expressão de seus genes parentais por meio de vias de siRNA endógenos. Pseudogenes surgem pela cópia de um gene ancestral. Alguns pseudogenes são transcritos e, dependendo do contexto genômico, podem produzir um RNA antissenso equivalente ao mRNA produzido pelo gene original. Um mRNA transcrito a partir do gene original (A) e um transcrito antissenso de um pseudogene correspondente (ΨA) podem então formar um RNA dupla-fita que será clivado pela dicer para originar um siRNA. siRNAs endógenos também podem ser produzidos por sequências invertidas duplicadas, tais como o exemplo representado aqui de uma duplicação invertida de um pseudogene (ΨA ΨA), à direita. A transcrição de ambas as cópias do pseudogene resulta em um longo RNA com repetições invertidas (em azul, encimado por setas), causando o dobramento do RNA em um grampo que será clivado pela dicer para originar um siRNA. Em ambos os casos, o siRNA endógeno é guiado pelo RISC para degradar e interagir com os transcritos de mRNA do gene original que ainda restam. Setas verdes indicam rearranjos de DNA. [Adaptada de Sasidharan R & Gerstein M (2008) *Nature* 453, 729 – 731. Com permissão de Macmillan Publishers Ltd.]

O envolvimento de ncRNAs longos na regulação de processos de desenvolvimento é ilustrado por uma análise abrangente (resolução de 5 pb) dos produtos transcricionais dos quatro agrupamentos do gene *HOX* humano. Embora existam apenas 39 genes *HOX*, o produto transcricional dos agrupamentos *HOX* inclui um total de 231 ncRNAs longos diferentes. Muitos destes são reguladores que atuam em *cis*, mas um deles, HOTAIR, foi identificado como regulador *trans* (ver Tabela 9.12).

Alguns dos RNAs funcionais, tais como XIST e AIR, não foram tão bem conservados ao longo da evolução. As sequências funcionais de mais rápida evolução no genoma humano incluem componentes de longos ncRNAs específicos de primatas que são fortemente expressos no cérebro. As implicações evolutivas de tais genes serão consideradas no Capítulo 10.

TABELA 9.12 Exemplos de longos RNAs regulatórios humanos

RNA	Tamanho	Localização do gene	Organização do gene	Função
XIST	19,3 kb	Xq13	6 éxons ocupando 32 kb	regulador da inativação do cromossomo X
TSIX	37 kb	Xq13	1 éxon	regulador antissenso de XIST
H19	2,3 kb	11p15	5 éxons ocupando 2,67 kb	envolvido no *imprinting* do agrupamento imprintado 11p15, associado com a síndrome de Beckwith-Wiedemann
KCNQTOT1 (= LIT1)	59,5 kb	11p15	1 éxon	regulador antissenso do agrupamento imprintado em 11p15
PEG3	1,8 kb[a]	19q13	número variável de éxons, mas até 9 éxons ocupando região de 25 kb	imprintado maternamente e com função reconhecida na supressão tumoral por ativação de p53
HOTAIR	2,2 kb	12q13	6 éxons ocupando 6,3 kb	regulador gênico em *trans*; embora seja parte de uma região regulatória do agrupamento *HOX-C* em 12q13, o RNA HOTAIR reprime a transcrição de uma região de 40 kb no agrupamento *HOX-D* do cromossomo 2q31

[a]Isoformas maiores.

9.4 DNA ALTAMENTE REPETITIVO: HETEROCROMATINA E REPETIÇÕES DE TRANSPOSONS

Os genes contêm algumas sequências de DNA repetitivo, incluindo DNA repetitivo codificante. Entretanto, a maioria das sequências de DNA altamente repetitivo ocorre fora dos genes. Algumas destas sequências estão presentes em certas regiões subcromossômicas sob a forma de grandes arranjos de repetições em *tandem*. Este tipo de DNA, conhecido como heterocromatina, permanece altamente condensado ao longo do ciclo celular e geralmente não apresenta genes.

Outras sequências de DNA altamente repetitivo estão dispersas ao longo do genoma humano e se originaram a partir de *transposição duplicativa* (ver Seção 9.1). Sequências como estas são descritas, às vezes, como *repetições de transposons* e correspondem a mais de 40% da sequência total de DNA do genoma humano. Além de residir em regiões extragênicas, elas são geralmente encontradas em íntrons e sequências não traduzidas e às vezes até mesmo em regiões codificantes.

A heterocromatina constitutiva é amplamente definida por longos arranjos de repetições de DNA de alto número de cópias em *tandem*

O DNA da heterocromatina constitutiva corresponde a 200 Mb ou 6,5% do genoma humano (ver Tabela 9.3). Ele engloba regiões de megabases nos centrômeros e regiões de DNA comparativamente mais curtas nos telômeros de todos os cromossomos. A maior parte do cromossomo Y e dos braços curtos dos cromossomos acrocêntricos (13, 14, 15, 21 e 22) consiste em heterocromatina. Além disso, há regiões heterocromáticas substanciais próximas do centrômero de certos cromossomos, notadamente nos cromossomos 1, 9, 16 e 19.

O DNA da heterocromatina constitutiva consiste principalmente em longos arranjos de sequências de DNA repetitivo de alto número de cópias em *tandem*, conhecidos como *DNA satélite* (**Tabela 9.13**). Arranjos menores de repetições em *tandem* são conhecidos como minissatélites e microssatélites. Grandes extensões de heterocromatina são geralmente compostas de um mosaico de diferentes sequências de DNA satélite, ocasional-

TABELA 9.13 Principais classes de DNA humano de alto número de cópias repetido em *tandem*

Classe[a]	Tamanho total do arranjo	Tamanho ou sequência da unidade de repetição	Principais localizações cromossômicas
DNA satélite[b]	geralmente centenas de quilobases		associado à heterocromatina
α (DNA alfoide)		171 pb	heterocromatina centromérica em todos os cromossomos
β (família *Sau*3A)		68 pb	principalmente na heterocromatina centromérica dos cromossomos 1, 9, 13, 14, 15, 21, 22 e Y
Satélite 1		25 – 48 pb (ricos em AT)	heterocromatina centromérica da maioria dos cromossomos e outras regiões heterocromáticas
Satélite 2		formas divergentes de ATTCC/GGAAT	maioria dos cromossomos (possivelmente, em todos)
Satélite 3		ATTCC/GGAAT	13p, 14p, 15p, 21p, 22p e heterocromatina em 1q, 9q e Yq12
DYZ19		125 pb	~400 kb em Yq11
DYZ2		rico em AT	Yq12; maior periodicidade que ~2470 pb
DNA minissatélite	0,1-20 kb		nos telômeros ou próximos a eles, em todos os cromossomos
Minissatélite telomérico		TTAGGG	todos os telômeros
Minissatélites hipervariáveis		9 – 64 pb	todos os cromossomos, associados à eucromatina, principalmente em regiões subteloméricas
DNA microssatélite	< 100 pb	geralmente, 1 – 4 pb	amplamente dispersos por todos os cromossomos

[a]A distinção entre satélite, minissatélite e microssatélite é feita com base no tamanho total do arranjo, e não no tamanho da unidade de repetição. [b]Arranjos de DNA satélite que consistem em unidades de repetição simples geralmente têm composições de bases que são radicalmente diferentes da média de 41% G+C (e, portanto, podem ser isolados por centrifugação em gradientes de densidade, onde podem ser diferenciados do DNA principal e aparecem como *bandas satélite* – daí seu nome).

mente interrompidas por repetições de transposons, sendo desprovidas de genes. As repetições de transposons também estão amplamente distribuídas na eucromatina e serão descritas a seguir.

A grande maioria do DNA heterocromático humano não foi sequenciado, devido a dificuldades técnicas em obter clones de DNA sobrepostos ordenados de maneira não ambígua. Portanto, apenas componentes curtos representativos de DNA centromérico foram sequenciados até o momento. A heterocromatina do cromossomo Y, no entanto, é uma exceção e está bem caracterizada. Existem diferentes organizações de DNA satélite, e a unidade de repetição pode ser uma sequência muito simples (com menos de 10 nucleotídeos) ou uma sequência moderadamente complexa que pode se estender por mais de 100 nucleotídeos (ver Tabela 9.13).

Em nível de sequência, o DNA satélite é geralmente pouco conservado entre as espécies. Sua função exata permanece desconhecida, embora alguns DNAs satélite humanos estejam implicados no funcionamento dos centrômeros cujo DNA consiste, em grande parte, de várias famílias de DNA satélite.

O centrômero é um domínio definido epigeneticamente. Sua função independe da sequência subjacente de DNA; em vez disso, depende da organização particular de sua cromatina, a qual, uma vez estabelecida, deve ser mantida de maneira estável ao longo de múltiplas divisões celulares. Dentre as várias famílias de DNA satélite associadas com os centrômeros humanos, apenas o satélite alfa é reconhecidamente encontrado em todos os centrômeros humanos, e sua unidade de repetição geralmente contém um sítio de ligação para uma proteína centromérica específica, a CENPB. Arranjos de satélite alfa clonados foram capazes de induzir a formação de centrômeros em células humanas, indicando que o satélite alfa deve ter um papel importante no funcionamento do centrômero.

O especializado DNA telomérico consiste em arranjos de tamanho médio de apenas algumas quilobases de extensão e constitui uma forma de *minissatélite de DNA*. Ao contrário do DNA satélite, o DNA minissatélite telomérico foi extraordinariamente conservado durante a evolução dos vertebrados e desempenha papel integral na função dos telômeros. Ele consiste em um arranjo em *tandem* de repetições do hexanucleotídeo TTAGGG que são sintetizadas pela ribonucleoproteína telomerase (ver Figura 2.13).

Repetições derivadas de transposons representam mais de 40% do genoma humano e surgiram principalmente por meio de intermediários de RNA

Praticamente todo o DNA repetitivo não codificante disperso no genoma humano é derivado de **transposons** (também chamados de *elementos transponíveis*), sequências móveis de DNA que podem migrar para diferentes regiões do genoma. Cerca de 45% do genoma pertencem a esta classe de sequência, mas muitas das sequências de DNA de cópia única também devem ter sido derivadas de cópias antigas de transposons que divergiram extensivamente ao longo de um período medido em escala evolutiva.

Em humanos e outros mamíferos há quatro classes principais de repetições de transposons, mas apenas uma minoria das repetições se transpõe ativamente. Conforme o método de transposição, as repetições podem ser organizadas em dois grupos:

- *Retrotransposons* (também abreviados como *retroposons*). Neste caso, o mecanismo de cópia lembra a maneira pela qual pseudogenes processados e retrogenes são gerados (ver Figura 9.12): uma transcriptase reversa converte um transcrito de RNA do retrotransposon em uma cópia de cDNA a qual então se integra em diferentes regiões do DNA genômico. Três classes principais de transposons de mamíferos utilizam este mecanismo de cópia e colagem: elementos nucleares intercalantes longos (LINEs), elementos nucleares intercalantes curtos (SINEs) e elementos semelhantes a retrovírus contendo longas repetições terminais.
- *Transposons de DNA*. Membros desta quarta classe de transposons migram diretamente sem copiar a sequência; esta é excisada e então reinserida em outro local do genoma (um mecanismo de corte e colagem).

Elementos transponíveis que podem se transpor de maneira independente são descritos como *autônomos*; aqueles que são dependentes, chamam-se *não autônomos* (**Figura 9.20**). Dentre as quatro classes de elementos transponíveis, os LINEs e SINEs são os mais predominantes; estas sequências serão descritas em maior detalhe a seguir. As outras duas classes são brevemente descritas aqui.

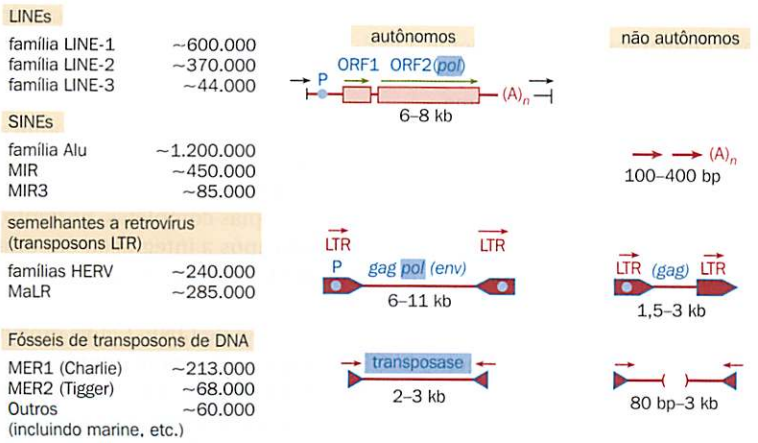


Figura 9.20 Famílias de transposons de mamíferos. Apenas uma pequena porção dos membros de qualquer das famílias de transposons aqui ilustradas parece ter capacidade de transposição; muitos perderam essa capacidade devido a mutações inativantes, e muitos são cópias curtas truncadas. Subclasses das quatro principais famílias são listadas, juntamente com seus respectivos tamanhos em pares de base. ORF, fase de leitura aberta. [Adaptada de International Human Genome Sequencing Consortium (2001) *Nature* 409, 860–921. Com permissão de Macmillan Publishers Ltd.]

Transposons LTR humanos

Transposons LTR incluem elementos semelhantes a retrovírus autônomos e não autônomos flanqueados por longas repetições terminais (LTRs) contendo elementos regulatórios transcricionais necessários. *Sequências retrovirais endógenas* contêm genes *gag* e *pol*, os quais codificam uma protease, uma transcriptase reversa, RNAse H e uma integrase. Eles estão aptos, portanto, a realizar sua transposição de maneira independente. Existem três classes principais de sequências retrovirais endógenas humanas (HERVs), com um número de cópias cumulativo de aproximadamente 240 mil, correspondendo a cerca de 4,6% do genoma humano (ver Figura 9.20).

Muitos HERVs são defeituosos, e sua transposição foi extremamente rara ao longo dos últimos milhões de anos. Entretanto, o grupo de sequências HERV-K, extremamente pequenas, apresenta conservação de genes retrovirais intactos, e alguns membros da subfamília HERV-K10 sofreram transposição relativamente recente durante a evolução. Elementos semelhantes a retrovírus não autônomos são desprovidos do gene *pol* e muitas vezes do gene *gag* (a sequência interna foi perdida por recombinação homóloga entre as LTRs flanqueadoras). A família MaLR de tais elementos responde por cerca de 4% do genoma humano.

Fósseis de transposons de DNA humanos

Transposons de DNA possuem repetições terminais invertidas e codificam uma transposase que regula sua transposição. Eles correspondem a cerca de 3% do genoma humano e podem ser agrupados em diferentes classes, as quais, por sua vez, são subdivididas em muitas famílias com origens independentes (ver o banco de dados de sequências de repetição Repbase em http://www.girinst.org/repbase/index.html). Existem duas famílias principais em humanos, MER1 e MER2, além de uma ampla variedade de famílias menos frequentes (ver Figura 9.20).

Praticamente todas as sequências de transposons de DNA residentes em humanos não são mais ativas; elas são, portanto, transposons fósseis. Os transposons de DNA tendem a ter vida curta dentro de uma espécie, ao contrário de alguns outros elementos transponíveis, tais como LINEs. Entretanto, alguns genes funcionais humanos parecem ter-se originado a partir de transposons de DNA, notadamente os genes que codificam as recombinases RAG1 e RAG2 e a principal proteína de ligação ao centrômero, a CENPB.

Alguns elementos humanos do tipo LINE-1 são transposons ativos e possibilitam a transposição de outros tipos de sequências de DNA

Sequências do tipo **LINE** (elemento nuclear intercalante longo) foram transposons muito bem-sucedidos. Eles têm uma história evolutiva relativamente longa, ocorrendo em outros mamíferos, inclusive em camundongos. Como transposons autônomos, eles podem produzir todas as moléculas necessárias para transposição, incluindo a transcriptase reversa, enzima essencial para este processo. LINEs humanos consistem em três famílias distantemente relacionadas: LINE-1, LINE-2 e LINE-3, coletivamente correspondendo a cerca de 20% de todo o genoma (ver Figura 9.20). Eles estão localizados predominantemente em regiões de eucromatina, preferencialmente em bandas G escuras (positivas para Giemsa), ricas em AT, de cromossomos metafásicos.

Dentre as três famílias de LINE humanas, LINE-1 (ou L1) é a única que continua a ter membros que se transpõem ativamente. LINE-1 é o elemento transponível humano mais importante e responde por uma fração de DNA genômico (17%) maior do que qualquer outra classe de sequências do genoma.

Elementos LINE-1 inteiros possuem mais de 6 kb de comprimento e codificam duas proteínas: uma proteína de ligação a RNA e uma proteína com funções de endonuclease e transcriptase reversa (**Figura 9.21A**). Extraordinariamente, um promotor interno está localizado na região 5′ não traduzida da sequência. Cópias completas, portanto, trazem consigo seu próprio promotor que pode ser utilizado após a integração em uma região permissiva do genoma. Após a tradução, o RNA do LINE-1 se complexa a proteínas codificadas por ele mesmo e se dirige ao núcleo.

Para a integração no DNA genômico, a endonuclease de LINE-1 cliva uma das duas fitas de DNA, deixando um grupamento 3′ OH livre que serve como ponto de início para a transcrição reversa a partir da extremidade 3′ do RNA de LINE. O sítio de clivagem preferencial da endonuclease é TTTT↓A; daí a preferência por integrar-se em *regiões ricas* em AT. Sequências de DNA ricas em AT são relativamente pobres em genes, e, portanto, como LINEs tendem a se integrar em regiões de DNA ricas em AT, eles impõem uma baixa carga mutacional, facilitando sua acomodação pelo hospedeiro. Durante a integração, a transcrição reversa geralmente não consegue alcançar a extremidade 5′, resultando em inserções truncadas, não funcionais. Em concordância a este fato, a maioria das repetições derivadas de LINEs é curta, com tamanho médio de 900 pb para todas as cópias de LINE-1, e apenas 1 entre 100 cópias é completa.

A maquinaria de LINE-1 é responsável pela maioria da transcrição reversa do genoma, permitindo a retrotransposição de SINEs não autônomos e também de cópias de mRNA, originando pseudogenes processados e retrogenes. Dentre as cerca de 6 mil sequências LINE-1 completas, aproximadamente 80 a 100 ainda são capazes de transposição e ocasionalmente causam doenças ao interromper a função de genes após sua inserção em importantes sequências conservadas.

Repetições Alu são os elementos mais numerosos do DNA humano e se originaram como cópias de moléculas de RNA 7SL

Os **SINEs** (elementos nucleares intercalantes curtos) são retrotransposons contendo 100-400 pb de comprimento. Eles foram muito bem-sucedidos na colonização de genomas de mamíferos, resultando em várias famílias de DNA intercalante, algumas com número de cópias extremamente alto. Ao contrário dos LINEs, os SINEs não codificam proteínas e são incapazes de transposição independente. Entretanto, SINEs e LINEs compartilham sequências em suas extremidades 3′, e os SINEs podem ser mobilizados por LINEs vizinhos. Ao parasitarem a maquinaria de transposição dos elementos LINE, os SINEs podem atingir um número de cópias bastante elevado.

A família Alu humana é a família de SINEs mais proeminente em termos de número de cópias e é a sequência mais abundante do genoma humano, ocorrendo em média mais do que uma vez a cada 3 kb. A repetição Alu completa possui cerca de 280 pb de extensão e consiste em duas repetições em *tandem*, cada uma delas com cerca de 120

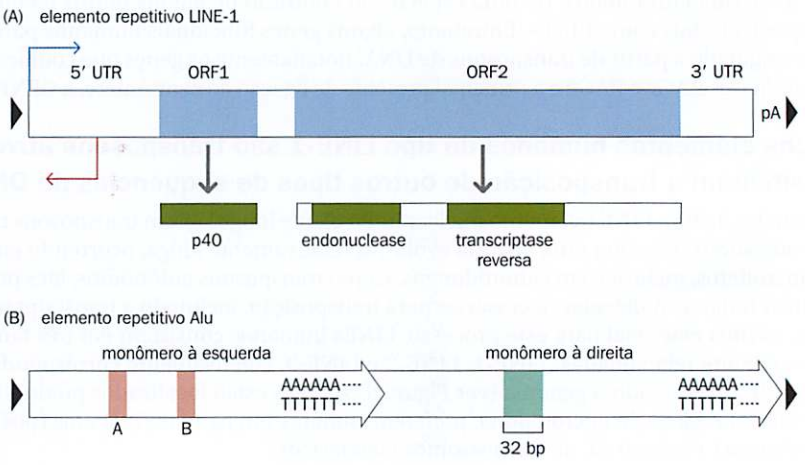


Figura 9.21 Os elementos repetitivos humanos LINE-1 e Alu. (A) O elemento LINE-1 contendo 6,1 kb possui duas fases abertas de leitura: ORF1, com 1 kb, codifica p40, uma proteína de ligação ao RNA com atividade de chaperona de ácidos nucleicos; a ORF2, com 4 kb, especifica uma proteína com atividades de endonuclease e transcriptase reversa. Um promotor interno bidirecional localiza-se dentro da região 5′ não traduzida (UTR). Na outra extremidade, há uma sequência A_n/T_n, geralmente descrita como a cauda poli(A) de 3′ (pA). A endonuclease de LINE-1 cliva uma das fitas do dúplex de DNA, preferencialmente na sequência TTTT↓A, e a transcriptase reversa utiliza a extremidade 3′-OH liberada para iniciar a síntese de cDNA. Novos sítios de inserção são flanqueados por pequenos sítios-alvo de duplicação de 2-20 pb (setas pretas nas extremidades). (B) Um dímero de Alu. Os dois monômeros possuem sequências semelhantes que terminam em uma sequência A_n/T_n, mas diferem em tamanho devido à inserção de um elemento de 32 pb no interior da repetição maior. Monômeros de Alu também existem no genoma humano, bem como várias cópias truncadas tanto de monômeros como de dímeros.

pb seguida por uma sequência A_n/T_n curta. As repetições em *tandem* são assimétricas: uma contém uma sequência interna de 32 pb que está ausente na outra (**Figura 9.21B**). Monômeros, contendo apenas uma das duas repetições em *tandem*, e várias versões truncadas de dímeros e monômeros também são comuns, resultando em uma média genômica geral de 230 pb.

Enquanto SINEs como as famílias MIR (repetição intercalante de mamíferos) são encontrados em uma grande variedade de mamíferos, a família Alu tem origem evolutiva relativamente recente e é encontrada apenas em primatas. Entretanto, subfamílias Alu de diferentes idades evolutivas podem ser identificadas. Nos últimos 5 milhões de anos, desde a divergência entre humanos e primatas africanos, apenas cerca de 5 mil cópias de repetições Alu sofreram transposição; as sequências Alu de maior mobilidade são os membros das subfamílias Y e S.

A exemplo de outros SINEs de mamíferos, as repetições Alu se originaram a partir de cópias de cDNA de pequenos RNAs transcritos pela RNA-polimerase III. Genes transcritos pela RNA-polimerase III geralmente têm promotores internos, e, portanto, cópias de cDNA de transcritos carregam consigo suas próprias sequências promotoras. Tanto as repetições Alu e, independentemente, as repetições B1 de camundongos, surgiram a partir de cópias de cDNA de RNA 7SL, o pequeno RNA que faz parte da partícula de reconhecimento de sinal, usando um mecanismo de retrotransposição como aquele apresentado na Figura 9.12. Outros SINEs, tais como a repetição B2 de camundongos, são cópias de sequências de tRNA retrotranspostas.

As repetições Alu têm conteúdo de GC relativamente alto e, embora dispersas predominantemente ao longo de regiões de eucromatina do genoma, estão preferencialmente localizadas em bandas R ricas em GC e genes, contrastando com a localização preferencial de LINEs em regiões de DNA ricas em AT. Entretanto, quando localizadas dentro de genes, como os elementos LINE-1, estas sequências estão confinadas a íntrons e regiões não traduzidas. Apesar da tendência em localizarem-se em regiões do DNA ricas em GC, repetições Alu recentemente transpostas apresentam preferência por regiões ricas em AT, e repetições Alu progressivamente antigas apresentam um forte viés em relação a regiões ricas em GC.

O viés na distribuição geral de repetições Alu em regiões ricas em GC e, consequentemente, regiões ricas em genes deve ser o resultado de uma forte pressão seletiva. Isso sugere que repetições Alu não são apenas parasitas do genoma, mas que são sequências úteis para as células que as possuem. Algumas sequências Alu são ativamente transcritas e podem ser recrutadas para desempenhar funções úteis. O gene *BCYRN1*, que codifica o RNA citoplasmático neuronal BC200, surgiu a partir de um monômero de Alu e é uma das poucas sequências Alu transcricionalmente ativas sob condições normais. Além disso, foi recentemente demonstrado que a repetição Alu atua como um repressor transcricional em *trans* durante a resposta celular ao choque térmico.

CONCLUSÃO

Neste capítulo, estudou-se a arquitetura do genoma humano. Cada célula humana contém muitas cópias de um genoma mitocondrial pequeno, circular, e apenas uma cópia do genoma nuclear, muito maior. Enquanto o genoma mitocondrial apresenta algumas semelhanças com o genoma compacto dos procariotos, o genoma nuclear humano é muito mais complexo em sua organização, com apenas 1,1% do genoma codificando proteínas e 95% dele correspondendo a sequências de DNA não conservadas e, geralmente, altamente repetitivas.

O sequenciamento do genoma humano revelou que, ao contrário do que se esperava, há relativamente poucos genes que codificam proteínas – 20.000 a 21.000, de acordo com estimativas mais recentes. Estes genes variam amplamente em tamanho e organização interna, com éxons geralmente separados por grandes íntrons, os quais frequentemente contêm sequências de DNA altamente repetitivo. A distribuição dos genes ao longo do genoma não é uniforme, com alguns genes relacionados estrutural e funcionalmente encontrados em agrupamentos, sugerindo que teriam surgido pela duplicação de genes individuais ou de segmentos maiores de DNA. Pseudogenes podem ser originados quando um gene é duplicado, e então uma das cópias acumula mutações deletérias, evitando sua expressão. Outros pseudogenes surgem quando um transcrito de RNA é reversamente transcrito e o cDNA gerado é reinserido no genoma.

A maior surpresa da era pós-genômica é o número e a variedade de RNAs não codificantes de proteínas transcritos a partir do genoma humano. Sabe-se agora que pelo

menos 85% do genoma eucromático é transcrito. Às já familiares moléculas de ncRNAs, que atuam na síntese proteica, juntaram-se outras que desempenham funções na regulação gênica, incluindo várias classes prolíficas de pequenos RNAs regulatórios e milhares de diferentes ncRNAs longos. Nossa visão tradicional do genoma está sendo radicalmente alterada.

No Capítulo 10, será descrito como o genoma humano se compara a outros genomas e como a evolução o moldou. Aspectos da expressão gênica humana são elaborados no Capítulo 11. No Capítulo 13, também será considerada a diversidade do genoma humano.

LEITURAS ADICIONAIS

Genoma mitocondrial humano

Anderson S, Bankier AT, Barrell BG et al. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.

Chen XJ & Butow RA (2005) The organization and inheritance of the mitochondrial genome. *Nat. Rev. Genet.* 6, 815–825.

Falkenberg M, Larsson NG & Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. *Annu. Rev. Biochem.* 76, 679–699.

MITOMAP: human mitochondrial genome database. http://www.mitomap.org

Wallace DC (2007) Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. *Annu. Rev. Biochem.* 76, 781–821.

Genoma nuclear humano

Clamp M, Fry B, Kamal M et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* 104, 19428–19433.

Ensembl human database. http://www.ensembl.org/Homo_sapiens/index.html

GeneCards human gene database. http://www.genecards.org

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.

Nature Collections: Human Genome Supplement, 1 June 2006 issue. [Uma compilação de artigos que analisa a sequência de cada cromossomo, com cópias dos artigos de 2001 e de 2004 que reportaram, respectivamente, o rascunho e o resultado final da sequência de eucromatina, disponível eletronicamente em http://www.nature.com/nature/supplements/collections/humangenome/]

NCBI Human Genome Resources. http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/

UCSC Genome Browser, Human (*Homo sapiens*) Genome Browser Gateway. http://genome.ucsc.edu/cgi-bin/hgGateway

Organização de genes que codificam proteínas

Adachi N & Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 109, 807–809.

Li YY, Yu H, Guo ZM et al. (2006) Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput. Biol.* 2, e74.

Sanna CR, Li W-H & Zhang L (2008) Overlapping genes in the human and mouse genomes. *BMC Genomics* 9, 169.

Soldà G, Suyama M, Pelucchi P et al. (2008) Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genomics* 9, 174.

Duplicação gênica, duplicação segmentar e variação do número de cópias

Bailey JA, Gu Z, Clark RA et al. (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007.

Conrad B & Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet.* 8, 17–35.

Kaessmann H, Vinckenbosch N & Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10, 19–31.

Linardopoulou EV, Williams EM, Fan Y et al. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437, 94–100.

Redon R, Ishikawa S, Fitch KR et al. (2006) Global variation in copy number in the human genome. *Nature* 444, 444–454.

Tuzun E, Sharp AJ, Bailey JA et al. (2005) Fine-scale structural variation of the human genome. *Nat. Genet.* 37, 727–732.

A complexidade do transcriptoma de mamíferos e a necessidade de redefinir genes na era pós-sequenciamento genômico

Gerstein MB, Bruce C, Rozowsky JS et al. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.* 17, 669–681.

Gingeras T (2007) Origin of phenotypes: genes and transcripts. *Genome Res.* 17, 682–690.

Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* 10, 833–844.

Kapranov P, Cheng J, Dike S et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

Revisões gerais sobre RNA não codificante

Amaral PP, Dinger ME, Mercer TR & Mattick JS (2008) The eukaryotic genome as an RNA machine. *Science* 319, 1787–1789.

Carninci P, Yasuda J & Hayashizaki Y (2008) Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol.* 20, 274–280.

Griffiths-Jones S (2007) Annotating non-coding RNA genes. *Annu. Rev. Genomics Hum. Genet.* 8, 279–298.

Marakova JA & Kramerov DA (2007) Non-coding RNAs. *Biochemistry (Moscow)* 72, 1161–1178.
Mattick JS (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.* 5, e1000459.
Prasanth KV & Spector DL (2007) Eukaryotic regulatory RNAs: an answer to the genome complexity conundrum. *Genes Dev.* 21, 11–42.

Pequenos RNAs nucleares e pequenos RNAs nucleolares

Kishore S & Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311, 230–232.
Matera AG, Terns RM & Terns MP (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* 8, 209–220.
Sahoo T, del Gaudio D, German JR et al. (2008) Prader–Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nuclear RNA cluster. *Nat. Genet.* 40, 719–721.

MicroRNAs e ncRNAs como reguladores do desenvolvimento

Bushati N & Cohen SM (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol.* 23, 175–205.
Chang T-C & Mendell JT (2007) microRNAs in vertebrate physiology and disease. *Annu. Rev. Genomics Hum. Genet.* 8, 215–239.
Makeyev EV & Maniatis T (2008) Multilevel regulation of gene expression by microRNAs. *Science* 319, 1789–1790.
Rinn JL, Kertesz M, Wang JK et al. (2007) Functional demarcation of active and silent chromatin domains in human *HOX* loci by non-coding RNAs. *Cell* 129, 1311–1323.
Stefani G & Slack F (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.* 9, 219–230.

piRNAs e siRNAs endógenos

Aravin AA, Sachidanandam R, Girard A et al. (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316, 744–747.
Girard A & Hannon GJ (2007) Conserved themes in small RNA-mediated transposon control. *Trends Cell Biol.* 18, 136–148.
Tam OH, Aravin AA, Stein P et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453, 534–538.
Watanabe T, Totoki Y, Toyoda A et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453, 539–543.

RNAs antissenso e longos RNAs não codificantes regulatórios

He Y, Vogelstein B, Velculescu VE et al. (2008) The antisense transcriptomes of human cells. *Science* 322, 1855–1858.
Khalil AM, Guttman M, Huarte M et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* 106, 11667–11672.
Ponting CP, Oliver PL & Reik W (2009) Evolution and function of long noncoding RNAs. *Cell* 136, 629–641.
Wilusz JE, Sunwoo H & Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* 23, 1494–1504.

RNAs associados a promotores e regiões terminais

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457, 1028–1042.

Pseudogenes e retrogenes

D'Errico L, Gadaleta G & Saccone C (2004) Pseudogenes in metazoa: origins and features. *Brief. Funct. Genomic. Proteomic.* 3, 157–167.
Duret L, Chureau C, Samain S et al. (2006) The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312, 1653–1655.
Sasidharan R & Gerstein M (2008) Protein fossils live on as RNA. *Nature* 453, 729–731.
Zhang D & Gerstein MB (2004) Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev.* 14, 328–335.
Zheng D & Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* 23, 219–224.

Heterocromatina e repetições baseadas em transposons

Choo KH, Vissel B, Nagy A et al. (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res.* 19, 1179–1182.
Faulkner GJ, Kimura Y, Daub CO et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* 41, 563–571.
Henikoff S, Ahmad K & Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–1102.
Mariner PD, Walters RD, Espinoza CA et al. (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* 29, 499–509.
Mills RE, Bennett EA, Iskow RC & Devine SE (2007) Which transposable elements are active in the human genome? *Trends Genet.* 23, 183–191.
Muotri AR, Marchetto MCN, Coufal NG & Gage FH (2007) The necessary junk: new functions for transposable elements. *Hum. Mol. Genet.* 16, R159–R167.
Repbase: database of repeat sequences. http://www.girinst.org/repbase/index.html
Wicker T, Sabot F, Hua-Van A et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–983.
Yang N & Kazazian HH Jr (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol.* 13, 763–771.