CAPÍTULO 10

# Identificação da Base Genética para Doenças Humanas

Este capítulo fornece uma visão geral de como os geneticistas estudam famílias e populações para identificar contribuições genéticas para uma doença. Independentemente de uma doença ser herdada em um padrão mendeliano reconhecível, tal como ilustrado no Capítulo 7, ou apenas ocorrer com uma frequência maior nos parentes dos indivíduos acometidos, como explorado no Capítulo 8, são as diferentes variantes genéticas e genômicas portadas pelos membros da família acometidos ou indivíduos acometidos na população que causam a doença de maneira direta ou influenciam sua suscetibilidade à doença. A investigação sobre o genoma tem fornecido aos geneticistas um catálogo de todos os genes humanos conhecidos, o conhecimento da sua localização e estrutura e uma lista crescente de dezenas de milhões de variantes na sequência de DNA encontradas entre indivíduos em diferentes populações. Como vimos nos capítulos anteriores, algumas dessas variantes são comuns, outras são raras, e ainda outras diferem em frequência entre diferentes grupos étnicos. Enquanto algumas variantes têm claramente consequências funcionais, outras certamente são neutras. Para a maioria, sua importância para a saúde humana e doenças é desconhecida.

No Capítulo 4, lidamos com o efeito da mutação, que altera um ou mais genes ou *loci* gerando alelos e polimorfismos variantes. E nos Capítulos 7 e 8 examinamos o papel dos fatores genéticos na patogenia de vários distúrbios mendelianos ou complexos. Neste capítulo, discutimos como os geneticistas abordam a descoberta de genes particulares implicados na doença e as variantes que eles contêm e que são subjacentes ou contribuem para doenças humanas, com foco em três abordagens.

- A primeira abordagem, **a análise de ligação**, é *baseada na família*. A análise de ligação obtém vantagens explícitas de heredogramas de famílias para acompanhar a herança de uma doença entre membros da família e para testar a co-hereditariedade consistente, repetida da doença com uma *região genômica particular* ou mesmo com uma *variante ou variantes específicas*, sempre que a doença é transmitida em uma família.
- A segunda abordagem, **a análise de associação**, é a *baseada na população*. A análise de associação não depende explicitamente de heredogramas, mas sim aproveita toda a história de uma população para procurar um aumento ou uma redução da frequência de um *alelo particular* ou *conjunto de alelos* em uma amostra de indivíduos acometidos coletada a partir da população, em comparação com um conjunto controle de pessoas não acometidas da mesma população. É particularmente útil para doenças complexas que não apresentam um padrão de herança mendeliana.
- A terceira abordagem envolve o **sequenciamento direto do genoma** dos indivíduos acometidos e de seus pais e/ou de outros indivíduos na família ou na população. Essa abordagem é particularmente útil para distúrbios mendelianos raros, nos quais a análise de ligação não é possível porque simplesmente não há famílias suficientes para fazer a análise de ligação ou porque o distúrbio é um letal genético que sempre resulta de mutações novas e nunca é herdado. Nestas situações, o sequenciamento do genoma (ou apenas dos éxons codificantes de cada gene, o **exoma**) de um indivíduo acometido e o peneiramento através dos bilhões resultantes (ou, no caso do exoma, dezenas de milhões) de bases de DNA têm sido usados com sucesso para encontrar o gene responsável pelo distúrbio. Esta nova abordagem aproveita a tecnologia recentemente desenvolvida que reduziu o custo do sequenciamento do DNA um milhão de vezes em comparação ao que era quando o genoma de referência original estava sendo preparado durante o Projeto Genoma Humano.

O uso de ligação, de associação e do sequenciamento para mapear e identificar os genes de doenças teve um enorme impacto sobre nossa compreensão da patogenia e fisiopatologia de muitas doenças. Com o tempo, o conhecimento das contribuições genéticas para a doença também irá sugerir novos métodos de prevenção, manejo e tratamento.

## BASE GENÉTICA PARA ANÁLISE DE LIGAÇÃO E ASSOCIAÇÃO

Uma característica fundamental da biologia humana é que cada geração se reproduz através da combinação de gametas haploides que contêm 23 cromossomos, resultantes da segregação independente e recombinação de cromossomos homólogos (Cap. 2). Para entender completamente os conceitos subjacentes à análise de ligação genética e os testes para a associação, é necessário revisar brevemente o comportamento de cromossomos e genes durante a

meiose à medida que são passados de uma geração para a seguinte. Parte desta informação repete o material clássico sobre gametogênese apresentado no Capítulo 2, ilustrando-o com novas informações que se tornaram disponíveis como resultado do Projeto Genoma Humano e suas aplicações para o estudo de variação humana.

## Segregação Independente e Recombinação Homóloga na Meiose

Durante a meiose I, os cromossomos homólogos alinham-se em pares ao longo do fuso meiótico. Os homólogos paternos e maternos trocam segmentos homólogos por meio do *crossing over* e da criação de novos cromossomos que são um "patchwork" que consiste em porções alternadas dos cromossomos da avó e dos cromossomos do avô (Fig. 2-15). Na família ilustrada na Figura 10-1, exemplos de cromossomos recombinados são mostrados na prole (geração II) do casal na geração I. Também é mostrado que o indivíduo na geração III herda um cromossomo materno que contém segmentos derivados de todos os quatro cromossomos de seus avós maternos. A criação de tais cromossomos patchwork enfatiza a noção de individualidade genética humana: cada cromossomo herdado por uma criança de um progenitor nunca é exatamente o mesmo que uma das duas cópias desse cromossomo no progenitor.

Embora nenhum dos dois cromossomos homólogos geralmente pareça idêntico sob o microscópio, eles diferem substancialmente no nível da sequência de DNA. Como discutido no Capítulo 4, estas diferenças na mesma posição (*locus*) em um par de cromossomos homólogos são

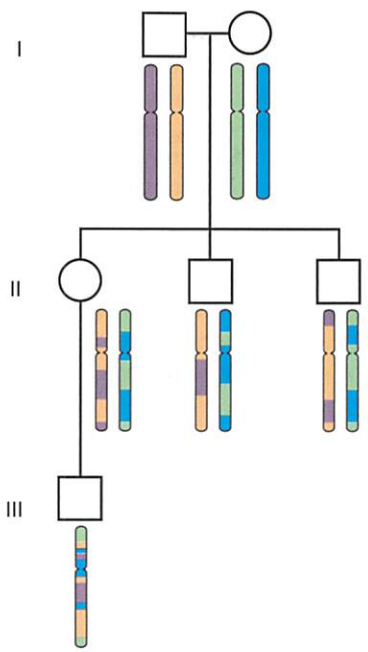



**Figura 10-1** Efeito de recombinação na origem de várias porções de um cromossomo. Devido ao *crossing over* na meiose, a cópia do cromossomo que o menino (geração III) herdou de sua mãe é um mosaico de segmentos de todas as quatro cópias daquele cromossomo de seus avós.

**alelos**. Alelos que são comuns (geralmente considerados como aqueles portados por aproximadamente 2% ou mais da população) constituem um **polimorfismo** e a análise da ligação em famílias (como iremos explorar mais tarde no capítulo) requer acompanhamento da herança de alelos específicos à medida que eles são transmitidos em uma família. As variantes alélicas em cromossomos homólogos possibilitam que os geneticistas tracem cada segmento de um cromossomo herdado por uma criança em particular para determinar se e onde os eventos de recombinação ocorreram ao longo dos cromossomos homólogos. Várias dezenas de milhões de marcadores genéticos estão disponíveis para servir como marcadores genéticos para esta finalidade. É uma trivialidade agora em genética humana dizer que é essencialmente sempre possível determinar com confiança, através de uma série de análises descritas neste capítulo, se um determinado alelo ou segmento do genoma em um paciente foi herdado de seu pai ou sua mãe. Este avanço - um produto singular do Projeto Genoma Humano - é uma característica essencial da análise genética para determinar a base genética precisa da doença.

### Alelos em *Loci* em Diferentes Cromossomos Segregam de Maneira Independente

Suponha que existem dois *loci* polimórficos, 1 e 2, em cromossomos diferentes, com alelos A e *a* no *locus* 1 e alelos *B* e *b* no *locus* 2 (Fig. 10-2). Suponha que um genótipo do indivíduo nesses *loci* é *Aa* e *Bb*; isto é, ela é heterozigota em ambos os *loci*, com alelos *A* e *B* herdados de seu pai e alelos *a* e *b* herdados de sua mãe. Os dois cromossomos diferentes irão se alinhar na placa metafásica na meiose I em uma de duas combinações de igual probabilidade. Depois da recombinação e da segregação cromossômica serem concluídas, haverá quatro possíveis combinações de alelos, *AB*, *ab*, *Ab* e *aB* em um gameta; cada combinação é tão provável de ocorrer como qualquer outra, um fenômeno conhecido como **segregação independente**. Pelo fato de os gametas *AB* conterem apenas os alelos derivados de seu pai, e os gametas *ab* apenas os alelos maternos, esses gametas são designados **parentais**. Em contrapartida, os gametas *Ab* ou *aB*, cada um contendo um alelo de origem paterna e um alelo de origem materna, são denominados gametas **não parentais**. Em média, a metade (50%) dos gametas será de parentais (*AB* ou *ab*) e 50% de não parentais (*Ab* ou *aB*).

### Alelos em *Loci* no Mesmo Cromossomo Segregam de Maneira Independente se Ocorrer pelo menos um *Crossover* entre Eles

Agora, suponha que um indivíduo é heterozigoto em dois *loci* 1 e 2, com os alelos *A* e *B* de origem paterna e *a* e *b* derivados maternalmente, mas os *loci* estão no mesmo cromossomo (Fig. 10-3). Os genes que residem no mesmo cromossomo são denominados **sintênicos** (literalmente, "no mesmo fio"), independentemente de quão próximos ou quão longe estejam naquele cromossomo.

Como esses alelos vão se comportar durante a meiose? Nós sabemos que entre um e quatro *crossovers* ocorrem entre cromossomos homólogos durante a meiose I quando

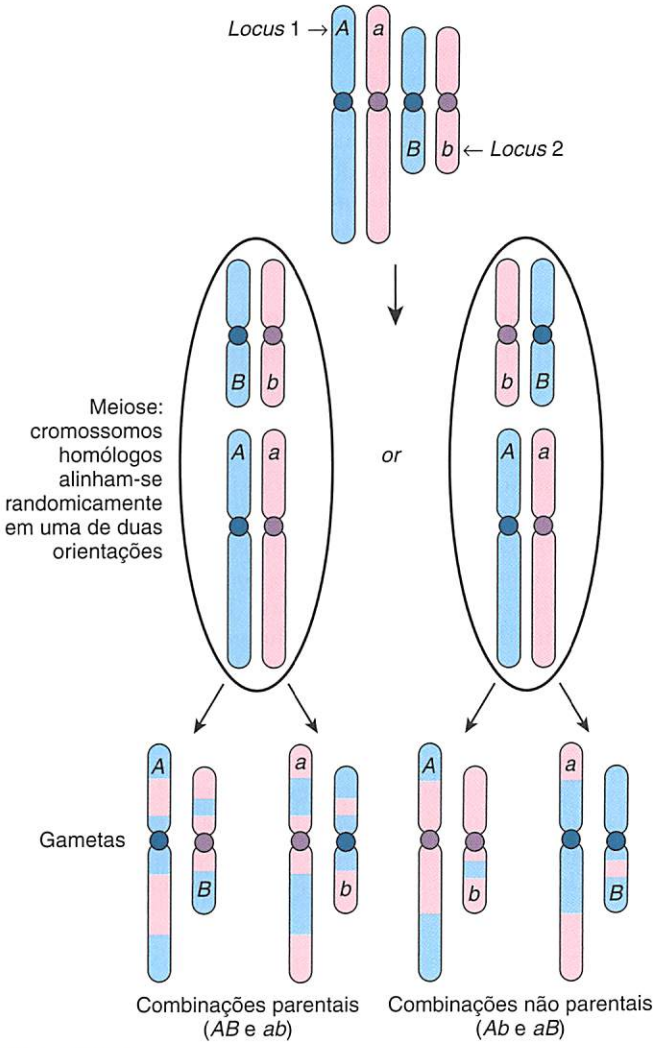



**Figura 10-2** **Segregação independente de alelos em dois *loci*, 1 e 2, quando estão localizados em cromossomos diferentes.** Suponha que os alelos A e B foram herdados de um dos progenitores, *a* e *b* do outro. Os dois cromossomos podem alinhar na placa de metáfase na meiose I, em uma de duas combinações igualmente prováveis, resultando em segregação independente dos alelos nestes dois cromossomos.

existem duas cromátides por cromossomo homólogo. Se nenhum *crossing over* ocorre dentro do segmento das cromátides entre os *loci* 1 e 2 (e ignorando tudo o que acontece em segmentos fora do intervalo entre esses *loci*), então os cromossomos que vemos nos gametas serão *AB* e *ab*, que são os mesmos dos cromossomos parentais originais; um cromossomo parental é, portanto, um cromossomo não recombinante. Se ocorrer *crossing over* pelo menos uma vez no segmento entre os *loci*, as cromátides resultantes podem ser ou não recombinantes ou *Ab* e *aB*, que não são as mesmas que os cromossomos parentais; esse cromossomo não parental é, portanto, um cromossomo **recombinante** (mostrado na Fig. 10-3). Uma, duas ou mais recombinações que ocorrem entre dois *loci* no estágio de quatro cromátides resultam em gametas que são 50% não recombinantes (parentais) e 50% recombinantes (não parentais), o que

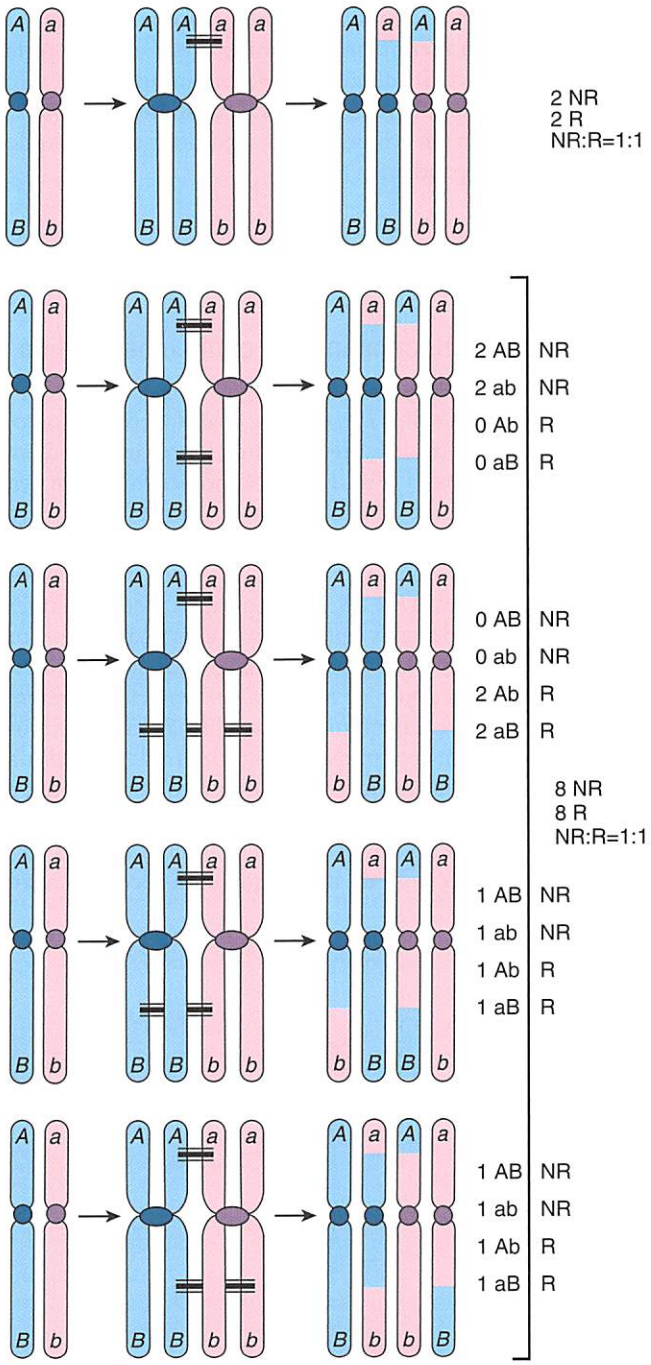



**Figura 10-3** *Crossing over* entre cromossomos homólogos (*linhas pretas horizontais*) na meiose é mostrado entre cromátides de dois cromossomos homólogos do lado *esquerdo*. Os *crossovers* resultam em novas combinações de alelos derivados da mãe e do pai nos cromossomos recombinantes presentes nos gametas, exibidos à *direita*. Se não ocorrer *crossing over* no intervalo entre os *loci* 1 e 2, apenas combinações de alelos parentais (não recombinantes), *AB* e *ab*, ocorrem na prole. Se um ou dois *crossovers* ocorrem no intervalo entre os *loci*, metade dos gametas conterá uma combinação de alelos não recombinantes e metade a combinação recombinante. O mesmo é verdadeiro se mais de dois *crossovers* ocorrerem entre os *loci* (não ilustrado aqui). NR, não recombinante; R, recombinante.

é precisamente nas mesmas proporções que se vê com segregação independente de alelos em *loci* em diferentes cromossomos. Assim, se dois *loci* sintênicos estão suficien-

temente afastados no mesmo cromossomo para assegurar que haverá pelo menos um *crossover* entre eles em toda meiose, a proporção de genótipos recombinantes e não recombinantes será, em média, de 1:1, exatamente como se os *loci* estivessem em cromossomos separados e segregando de maneira independente.

## Frequência de Recombinação e Distância do Mapa

### Frequência de Recombinação como uma Medida de Distância entre *Loci*

Agora suponha que dois *loci* estão no mesmo cromossomo, mas estão ou muito distantes, ou muito próximos, ou em algum ponto no meio (Fig. 10-4). Como acabamos de ver, quando os loci estão muito distantes (Fig. 10-4A), pelo menos um *crossover* ocorrerá no segmento do cromossomo entre os *loci* 1 e 2, e haverá gametas de ambos os genótipos não recombinantes *AB* e *ab* e os genótipos recombinantes *Ab* e *aB*, em proporções iguais (em média) na prole. Por outro lado, se dois *loci* estiverem muito próximos no mesmo cromossomo a ponto de os *crossovers nunca* ocorrerem entre eles, não haverá nenhuma recombinação; os genótipos não recombinantes (cromossomos parentais *AB* e *ab* na Fig. 10-4B) são transmitidos em conjunto o tempo todo, e a frequência dos genótipos recombinantes *Ab* e *aB* será 0. Entre esses dois extremos está a situação em que dois *loci* estão longe o suficiente de modo que uma recombinação entre os *loci* ocorre em algumas meioses, mas não em outras (Fig. 10-4C). Nessa situação, observam-se combinações não recombinantes de alelos na prole, quando nenhum cruzamento ocorreu e combinações recombinantes, quando uma recombinação ocorreu, mas a frequência de cromossomos recombinantes nos dois *loci* ficará entre 0% e 50%. O ponto crucial é que *quanto mais próximos dois* loci *estiverem, menor será a frequência de recombinação e menor o número de genótipos recombinantes observados na prole.*

### A Detecção de Eventos de Recombinação Requer Heterozigosidade e Conhecimento de Fase

Detectar os eventos de recombinação entre *loci* requer que (1) um progenitor seja heterozigoto (**informativo**) em ambos os *loci* e que (2) saibamos qual alelo no *locus* 1 está no mesmo cromossomo que o alelo no *locus* 2. Em um indivíduo que é heterozigoto em dois *loci* sintênicos, um com alelos A e *a*, o outro com *B* e *b*, o alelo encontrado no primeiro *locus* está no mesmo cromossomo com que o alelo no segundo *locus* define o que é chamado de **fase** (Fig. 10-5). Diz-se que o conjunto de alelos no mesmo homólogo (*A* e *B*, ou *a* e *b*) está em **acoplamento** (ou *cis*) e forma o que é conhecido como um **haplótipo** (Capítulos 7 e 8). Em contrapartida, alelos em diferentes homólogos (*A* e *b*, ou *a* e *B*) estão em **repulsão** (ou *trans*) (Fig. 10-5).

A Figura 10-6 mostra um heredograma de uma família com múltiplos indivíduos acometidos pela **retinite pigmentar** (RP) autossômica dominante, uma doença degenerativa da retina que causa cegueira progressiva em associação

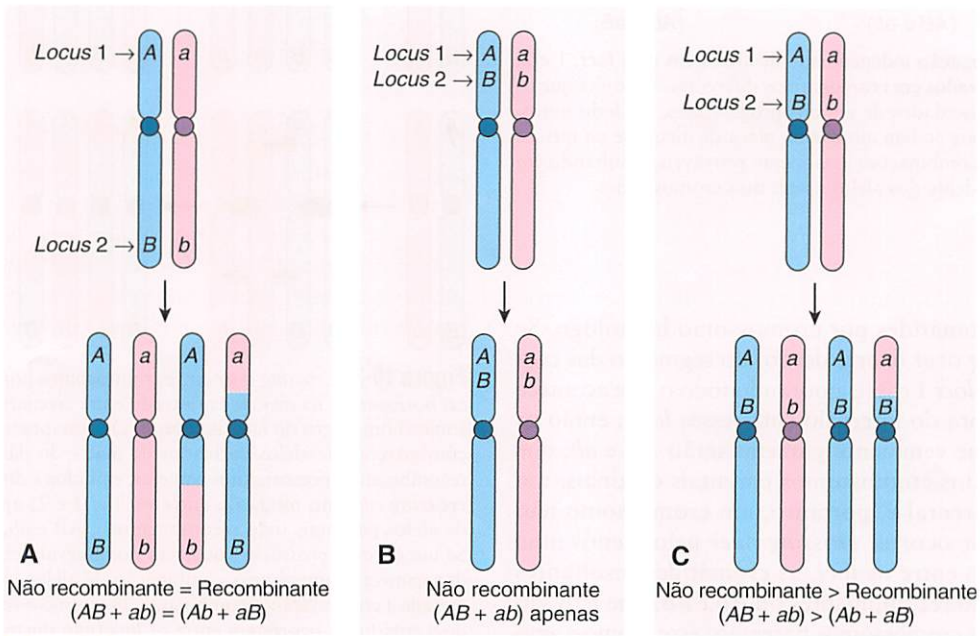
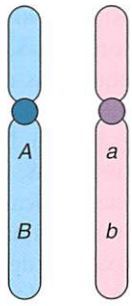



**Figura 10-4** Segregação de alelos em dois *loci*, 1 e 2, quando eles estão localizados no mesmo cromossomo. A, Os *loci* estão distantes e pelo menos um *crossover* entre eles tem probabilidade de ocorrer em cada meiose. B, Os *loci* estão tão próximos que o *crossing over* entre eles não é observado, independente da presença de *crossovers* em outra parte do cromossomo. C, Os *loci* estão próximos no mesmo cromossomo, mas suficientemente afastados para que ocorra *crossing over* no intervalo entre os dois *loci* apenas em algumas meioses, mas não em muitas outras.




Em acoplamento: *A* e *B* *a* e *b*
Em repulsão: *a* e *B* e *A* e *b*

**Figura 10-5** Possíveis fases de alelos *A* e *a*, e alelos *B* e *b*.

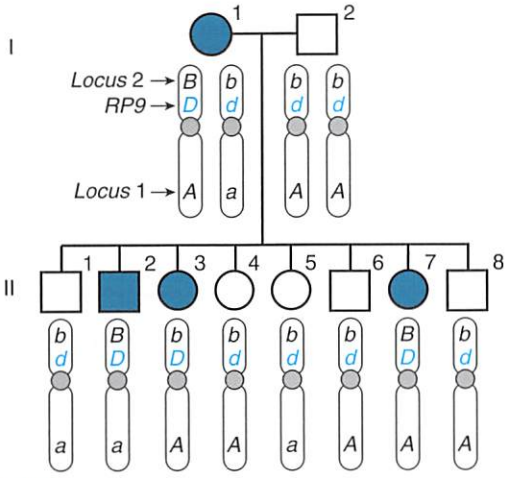



**Figura 10-6** **Co-hereditariedade do gene para uma forma autossômica dominante de retinite pigmentosa (RP), com *locus* marcador 2, e não com locus marcador 1.** Apenas a contribuição da mãe para os genótipos do filho é mostrada. A mãe (I-1) é acometida com esta doença dominante e é heterozigota no *locus RP9* (*Dd*), bem como nos *loci* 1 e 2. Ela carrega os alelos *A* e *B* no mesmo cromossomo que o alelo mutante *RP9* (D). O pai não acometido é homozigoto normal (*dd*) no *locus RP9*, bem como nos dois *loci* do marcador (*AA* e *bb*); suas contribuições para a prole não são consideradas. Dois dos três filhos acometidos herdaram o alelo *B* no *locus* 2 de sua mãe, enquanto o indivíduo II-3 herdou o alelo *b*. Os cinco filhos não acometidos também herdaram o alelo *b*. Assim, sete dos oito filhos são não recombinantes entre o *locus RP9* e o *locus* 2. No entanto, os indivíduos II-2, II-4, II-6 e II-8 são recombinantes para *RP9* e o *locus* 1, indicando que o *crossover* meiótico ocorreu entre esses dois *loci*.

com pigmentação anormal da retina. Como mostrado, o indivíduo I-1 é heterozigoto tanto no *locus* marcador 1 (com alelos *A* e *a*) como no *locus* marcador 2 (com alelos *B* e *b*), bem como heterozigoto para o distúrbio (*D* é o alelo da doença dominante, *d* é o alelo normal recessivo). Os alelos A-D-B formam um haplótipo, e *a-d-b* o outro. Como sabemos que seu esposo é homozigoto em todos os três *loci* e só pode repassar os alelos *a*, *b* e *d*, podemos facilmente determinar quais alelos as crianças receberam de sua mãe e assim traçar a herança de seu alelo causador de RP ou seu alelo normal naquele *locus*, bem como os alelos em ambos os *loci* marcadores em seus filhos. Uma inspeção rigorosa da Figura 10-6 possibilita determinar se cada criança herdou um haplótipo recombinante ou não recombinante da mãe.

No entanto, se a mãe (I-1) for homozigota *bb* no *locus* 2, então todos os filhos herdariam um alelo *b* materno, independente de terem recebido um alelo *D* mutante ou *d* normal no *locus RP9*. Por ela não ser informativa no *locus* 2 nesse caso, seria impossível determinar se a recombinação ocorreu. Da mesma maneira, se a informação fornecida para a família na Figura 10-6 fosse simplesmente que I-1 era heterozigota, *Bb*, no *locus* 2 e heterozigota em *locus* 2 para uma forma autossômica dominante de RP, mas a fase não era conhecida, não seria possível determinar quais de seus filhos eram não recombinantes entre o *locus RP9* e o *locus* 2 e quais de seus filhos eram recombinantes. Assim a determinação de quem é ou não é um recombinante requer que saibamos se o alelo *B* ou *b* no *locus* 2 estava no mesmo cromossomo que o alelo mutante *D* para RP no indivíduo I-1 (Fig. 10-6).

## Ligação e Frequência de Recombinação

**Ligação** é o termo utilizado para descrever um afastamento da segregação independente dos dois *loci*, ou, em outras palavras, a tendência dos alelos nos *loci* que estão muito próximos no mesmo cromossomo a serem transmitidos juntos, como uma unidade intacta, através da meiose. A análise da ligação depende da determinação da frequência de recombinação como uma medida da proximidade entre si dos dois *loci* em um cromossomo. Uma notação comum para a frequência de recombinação (como uma proporção, não uma porcentagem) é a letra grega teta, $\theta$, onde $\theta$ varia de 0 (nenhuma recombinação) a 0,5 (segregação independente). Se dois *loci* estiverem muito próximos a ponto de $\theta = 0$ entre eles (como na Fig. 10-4B), diz-se que eles estão **completamente ligados**; se eles estiverem tão distantes que $\theta = 0{,}5$ (como na Fig. 10-4A), eles estão segregando de maneira independente e são **não ligados**. Entre esses dois extremos existem vários graus de ligação.

## Mapas Genéticos e Mapas Físicos

A **distância no mapa** entre dois *loci* é um conceito *teórico* que se baseia nos dados *reais* – a extensão de recombinação observada, $\theta$, entre os *loci*. A distância no mapa é medida em unidades chamadas de **centimorgans** (cM), definidos como o comprimento genético sobre o qual, em média, um *crossover* ocorre em 1% das meioses. (O centimorgan é 1/100 de um "Morgan", em homenagem a Thomas Hunt Morgan, que observou pela primeira vez a recombinação genética na mosca de frutas, a *Drosophila*. Portanto, uma fração de 1% (ou seja, $\theta = 0{,}01$) traduz-se aproximadamente em uma distância do mapa de 1 cM. Como discutimos anteriormente neste capítulo, a frequência de recombinação entre dois loci aumenta proporcionalmente com a distância entre dois *loci* apenas até certo ponto porque, após os marcadores estarem distantes o suficiente para que pelo menos uma recombinação sempre ocorra, a frequência de recombinação observada será igual a 50% ($\theta = 0{,}5$), não importa o quão distantes fisicamente estejam os dois *loci*.

Para medir com precisão a verdadeira distância do mapa genético entre dois *loci* amplamente espaçados, portanto, é necessário que se utilize marcadores espaçados a

distâncias genéticas curtas (1 cM ou menos) no intervalo entre estes dois *loci*, e em seguida, adicionar os valores de θ entre os marcadores intercalares, porque os valores de θ entre pares de marcadores estreitamente ligados serão boas aproximações das distâncias genéticas entre eles. Usando essa abordagem, o comprimento genético de um genoma humano inteiro foi medido e, curiosamente, verificou-se que diferiam entre os sexos. Quando medido na meiose do sexo feminino, o comprimento genético do genoma humano é aproximadamente 60% maior (≈ 4.596 cM) do que quando ele é medido na meiose masculina (2.868 cM), e essa diferença entre os sexos é consistente e uniforme em cada autossomo. O comprimento genético médio sexual de todo o genoma humano haploide, que é estimado como contendo aproximadamente 3,3 bilhões de pares de base de DNA, ou ≈ 3.300 Mb (Cap. 2), é de 3.790 cM, para uma média de aproximadamente 1,15 cM/Mb. A razão para a recombinação aumentada observada por unidade de comprimento de DNA em mulheres em comparação com o sexo masculino é desconhecida, embora se possa especular que tem a ver com o aumento da oportunidade de *crossing over* promovido pelos muitos anos em que os precursores do gameta feminino permanecem na meiose I antes da ovulação (Cap. 2).

As medidas de recombinação de pares de recombinação entre os marcadores genéticos separados por 1 Mb ou mais fornecem uma razão quase constante entre distância genética e distância física de aproximadamente 1 cM/Mb. No entanto, quando a recombinação é medida com resolução muito maior, tais como entre marcadores espaçados com menos de 100 kb, a recombinação por unidade de comprimento torna-se não uniforme e pode variar em mais de quatro ordens de magnitude (0,01 a 100 cM/Mb). Quando visualizados na escala de algumas dezenas de pares de quilobases de DNA, a relação linear aparente entre a distância física em pares de base e recombinação entre marcadores polimórficos localizados a milhões de pares de base de DNA de distância é, na verdade, resultado de uma média dos chamados **pontos quentes de recombinação** intercalados entre as regiões de pouca ou nenhuma recombinação. Os pontos quentes ocupam apenas aproximadamente 6% da sequência no genoma e ainda são responsáveis por aproximadamente 60% de toda a recombinação meiótica no genoma humano. A base biológica para estes pontos quentes de recombinação é desconhecida. O impacto dessa não uniformidade de recombinação em alta resolução é discutida a seguir, quando abordamos o fenômeno de desequilíbrio de ligação.

## Desequilíbrio de Ligação

Em geral, o desequilíbrio de ligação é o caso em que os dois alelos em dois *loci* não vão apresentar qualquer fase preferida na população se os *loci* estiverem ligados, mas a uma distância de 0,1 cM a 1 cM ou mais. Por exemplo, suponha que os *loci* 1 e 2 estão a 1 cM de distância. Além disso, suponha que o alelo *A* está presente em 50% dos cromossomos em uma população e o alelo *a* nos outros 50% dos cromossomos, enquanto que no *locus* 2, um alelo *S* de suscetibilidade à doença está presente em 10% dos cromossomos e o alelo de proteção *s* está em 90% (Fig. 10-7). Pelo fato de a frequência do haplótipo A-S, freq(*A-S*), ser simplesmente o produto das frequências de dois alelos – freq.(*A*) x freq(*S*) = 0,5 x 0,1 = 0,05, diz-se que os alelos estão em **equilíbrio de ligação** (Fig. 10-7A). Isto é, as frequências dos quatro haplótipos possíveis, *A-S*, *A-s*, *a-S* e *a-s* decorrem diretamente das frequências alélicas de A, *a*, *S* e *s*.

No entanto, ao examinarmos haplótipos que envolvem *loci* que estão muito próximos, descobrimos que saber as frequências alélicas para esses *loci* individualmente *não* possibilita prever as quatro frequências de haplótipos. A frequência de qualquer um dos haplótipos, freq(*A-S*) por

**Equilíbrio de ligação: Frequências de haplótipo são como esperado de frequências alélicas**

A

| | | Frequências alélicas no *locus* 2 | |
|---|---|---|---|
| | | freq(S) = 0,1 | freq(s) = 0,9 |
| Frequências alélicas no *locus* 1 | freq(A) = 0,5 | Haplótipo *A-S* freq(*A-S*) = 0,05 | Haplótipo *A-s* freq(*A-s*) = 0,45 |
| | freq(*a*) = 0,5 | Haplótipo *a-S* freq(*a-S*) = 0,05 | Haplótipo *a-s* freq(*a-s*) = 0,45 |

**Desequilíbrio de ligação: Frequências de haplótipo divergem do que é esperado de frequências alélicas**

B

| | | Frequências alélicas no *locus* 2 | |
|---|---|---|---|
| | | freq(*S*) = 0,1 | freq(*s*) = 0,9 |
| Frequências alélicas no *locus* 1 | freq(*A*) = 0,5 | Haplótipo *A-S* freq(*A-S*) = 0 | Haplótipo *A-s* freq(*A-s*) = 0,5 |
| | freq(*a*) = 0,5 | Haplótipo *a-S* freq(*a-S*) = 0,1 | Haplótipo *a-s* freq(*a-s*) = 0,4 |

**Desequilíbrio de ligação parcial: Frequências de haplótipo são mais raras do que o esperado de frequências alélicas**

C

| | | Frequências alélicas no *locus* 2 | |
|---|---|---|---|
| | | freq(*S*) = 0,1 | freq(s) = 0,9 |
| Frequências alélicas no *locus* 1 | freq(*A*) = 0,5 | Haplótipo *A-S* freq(*A-S*) = 0,01 | Haplótipo *A-s* freq(*A-s*) = 0,49 |
| | freq(*a*) = 0,5 | Haplótipo *a-S* freq(*a-S*) = 0,09 | Haplótipo *a-s* freq(*a-s*) = 0,41 |

**Figura 10-7** **Tabelas demonstrando como as mesmas frequências alélicas podem resultar em diferentes frequências de haplótipos indicativos de equilíbrio de ligação, forte desequilíbrio de ligação ou desequilíbrio de ligação parcial.** A, Sob equilíbrio de ligação, frequências do haplótipo são, como esperado, do produto das frequências de alelo relevantes. B, *Loci* 1 e 2 estão localizados muito próximos um do outro, e alelos nesses *loci* apresentam forte desequilíbrio de ligação. O haplótipo *A-S* está ausente e *a-s* é menos frequente (0,4 em vez de 0,45) comparado com o que é esperado de frequências alélicas. C, Alelos em *loci* 1 e 2 mostram desequilíbrio de ligação parcial. Haplótipos, *A-S* e *a-s* estão subrrepresentados em comparação com o que se espera de frequências alélicas. Observe que as frequências alélicas para *A* e *a* no *locus* 1 e para *S* e *s* no *locus* 2 são as mesmas em todas as três tabelas; é a maneira como os alelos são distribuídos nos haplótipos, mostrada nas quatro células centrais da tabela, que diferem.

exemplo, pode *não* ser igual ao produto das frequências dos alelos únicos que compõem aquele haplótipo; nesta situação, freq (*A-S*) ≠ freq (*A*) x freq (*S*) e assim se diz que os alelos estão em **desequilíbrio de ligação** (DL). O desvio ("delta") entre as frequências de haplótipos esperadas e reais é chamado de D e é fornecido por:

$$D = \text{freq}(A\text{-}S) \times \text{freq}(a\text{-}s) - \text{freq}(A\text{-}s) \times \text{freq}(a\text{-}S)$$

$D \neq 0$ é equivalente a dizer que os alelos estão em DL, enquanto $D = 0$ significa que os alelos estão em equilíbrio de ligação.

Exemplos de DL são ilustrados nas Figuras 10-7B e 10-7C. Suponha que se descobre que *todos* os cromossomos portadores do alelo *S* também têm o alelo *a*, enquanto nenhum tem um alelo *A* (Fig. 10-7B). Então diz-se que o alelo *S* e o alelo *a* estão em DL completo. Como um segundo exemplo, suponha que o haplótipo *A-S* está presente em apenas 1% de cromossomos na população (Fig. 10-7C). O haplótipo *A-S* tem uma frequência muito abaixo do que seria de se esperar com base nas frequências de alelos *A* e *S* na população como um todo, e $D < 0$, enquanto que o haplótipo *a-S* tem uma frequência muito maior do que o esperado e $D > 0$. Em outras palavras, os cromossomos que portam o alelo de susceptibilidade são enriquecidos para o alelo *a* às custas do alelo *A*, em comparação com cromossomos que portam o alelo de proteção *s*. Note, porém, que as frequências alélicas individuais permanecem inalteradas; apenas o que difere é como são distribuídos em haplótipos e isso é o que determina se há DL.

## Desequilíbrio de Ligação tem tanto Causas Biológicas quanto Históricas

O que causa DL? Quando um alelo da doença entra pela primeira vez na população (por mutação ou por imigração de um fundador portador do alelo da doença), o conjunto particular de alelos em *loci* polimórficos ligados ao (isto é, sintênicos com) *locus* da doença constitui um **haplótipo que contém doença** em que o alelo da doença está localizado (Fig. 10-8). O grau ao qual este haplótipo contendo doença original irá persistir ao longo do tempo depende em parte da probabilidade de que a recombinação move o alelo da doença *fora* do haplótipo original e *sobre* cromossomos com diferentes conjuntos de alelos nesses *loci* ligados. A velocidade com que a recombinação vai passar o alelo da doença para um novo haplótipo depende de um número de fatores:

- O número de gerações (e, portanto, o número de oportunidades para recombinação) desde a primeira aparição da mutação.

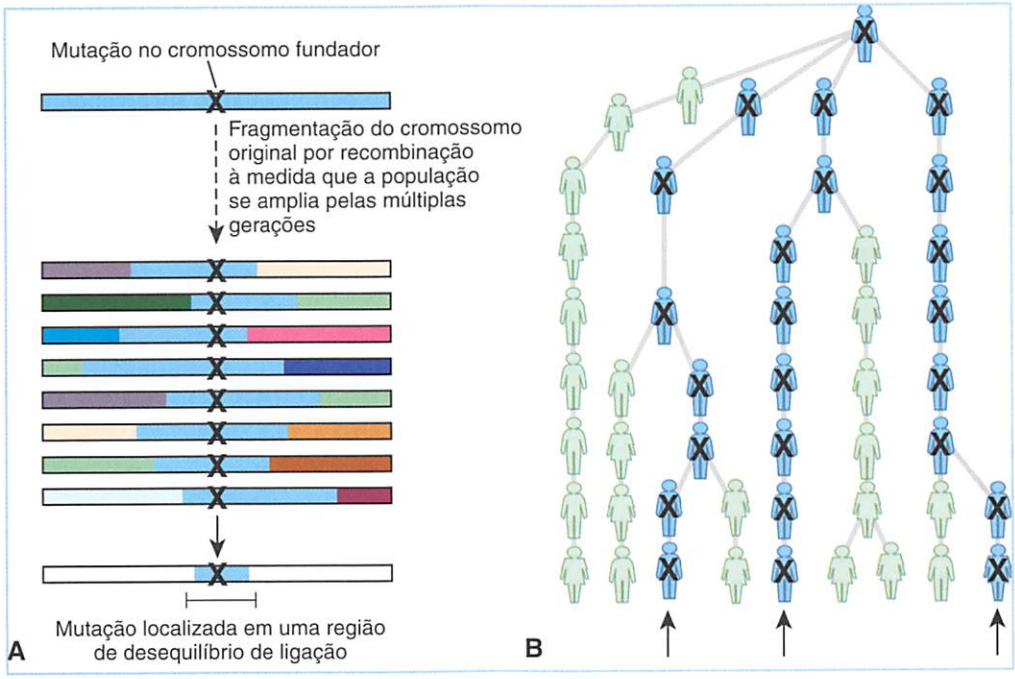



**Figura 10-8** Com cada geração, a recombinação meiótica troca os alelos que estavam inicialmente presentes em *loci* polimórficos em um cromossomo em que uma mutação associada à doença surgiu ( ) para outros alelos presentes no cromossomo homólogo. Ao longo de muitas gerações, os únicos alelos que permaneceram em fase de acoplamento com a mutação são aqueles nos *loci* tão perto do *locus* mutante que a recombinação entre os *loci* é muito rara. Estes alelos estão em desequilíbrio de ligação com a mutação e constituem um haplótipo associado à doença. **B**, Os indivíduos acometidos na geração atual (*setas*) são portadores da mutação (X) no desequilíbrio de ligação com o haplótipo associado à doença (*indivíduos em azul*). Dependendo da idade da mutação e de outros fatores genéticos da população, um haplótipo associado à doença geralmente se estende por uma região de DNA de alguns kb a algumas centenas de kb. *Veja Fontes & Agradecimentos.*

- A frequência de recombinação por geração entre os *loci*. Quanto menor for o valor de θ, maior a oportunidade de que o haplótipo contendo a doença persista intacto.
- Processos de seleção natural para ou contra determinados haplótipos. Se uma combinação de haplótipos sofre seleção positiva (e é, portanto, preferencialmente passada adiante) ou experimenta seleção negativa (e por isso é menos facilmente transmitida), será ou super-representada ou sub-representada nessa população.

### Medição do Desequilíbrio de Ligação

Embora conceitualmente valiosa, a discrepância, D, entre as frequências esperadas e observadas de haplótipos não é uma boa maneira de quantificar o DL, porque varia não só com o grau de DL, mas também com as frequências do alelo em si. Para quantificar diferentes graus de DL, consequentemente, os geneticistas frequentemente usam uma medida derivada da D, chamada de D' (Quadro). D' é concebido para variar de 0, indicando equilíbrio de ligação, até um máximo de ± 1, indicando DL muito forte. Pelo fato de o DL ser resultado não apenas da distância genética, mas também da quantidade de tempo durante o qual a recombinação teve uma chance de ocorrer e os possíveis efeitos de seleção para ou contra determinados haplótipos, diferentes populações vivendo em diferentes ambientes e com diferentes histórias podem ter diferentes valores de D' entre os mesmos dois alelos no mesmo *locus* do genoma.

$$D' = D/F$$

$$\text{Em que } D = \text{freq}(A\text{-}S) \times \text{freq}(a\text{-}s) - \text{freq}(A\text{-}s) \times \text{freq}(a\text{-}S)$$

e F é um fator de correção que ajuda a explicar as frequências alélicas.

O valor de F depende de se D em si é um número positivo ou um número negativo.

$$F = \text{o menor de freq}(A) \times \text{freq.(s) ou freq}(a) \times \text{freq}(S) \text{ se } D > 0$$

$$F = \text{o menor de freq}(A) \times \text{freq.(}S\text{) ou freq}(a) \times \text{freq}(s) \text{ se } D < 0$$

### Agrupamentos de Alelos Formam Blocos Definidos por Desequilíbrio de Ligação

A análise de medições por pares de D' para variantes vizinhas, particularmente polimorfismos de nucleotídeo único (SNPs) em todo o genoma revela uma arquitetura genética complexa para o DL. Os SNPs contíguos podem ser agrupados em aglomerados (*clusters*) de tamanho variável em que os SNPs em qualquer aglomerado apresentam níveis altos de DL uns com os outros, mas não com os SNPs fora desse agrupamento (Fig. 10-9). Por exemplo, os nove *loci* polimórficos no agrupamento 1 (Fig. 10-9A), cada um consistindo de dois alelos, têm o potencial de gerar $2^9 = 512$ haplótipos diferentes; no entanto, *apenas cinco haplótipos constituem 98% de todos os haplótipos observados*. Os valores absolutos de |D'| entre SNPs dentro do agrupamento estão bem acima de 0,8. Os agrupamentos de *loci* com alelos em DL alto em segmentos de apenas alguns pares de quilobases até algumas dezenas de pares de quilobase são denominados **blocos de DL**.

O tamanho de um bloco de DL que compreende alelos em um determinado conjunto de *loci* polimórficos não é idêntico em todas as populações. As populações africanas têm blocos menores, com média de 7,3 kb por bloco em todo o genoma, comparados com 16,3 kb em europeus; os tamanhos dos blocos chineses e japoneses são comparáveis entre si e são intermediários, atingindo uma média de 13,2 kb. Essa diferença no tamanho do bloco é quase certamente resultado do menor número de gerações desde a fundação das populações não africanas em comparação com as populações na África, desse modo limitando o tempo em que houve oportunidade de recombinação para quebrar regiões do DL.

Existe uma base biológica para os blocos de DL ou eles são simplesmente fenômenos genéticos que refletem uma história (e genoma) humana? Parece que a biologia realmente contribui para a estrutura do bloco de DL em que as fronteiras entre os blocos de DL frequentemente coincidem com pontos quentes de recombinação meiótica, discutidos anteriormente (Fig. 10-9C). Esses *hot spots* de recombinação quebrariam quaisquer haplótipos em dois haplótipos mais curtos mais rapidamente que a média, resultando em *equilíbrio* de ligação entre SNPs de um lado e do outro lado do *hot spot*. A correlação não é de maneira alguma exata, e muitas fronteiras aparentes entre blocos de DL não estão localizadas ao longo de *hot spots* de recombinação evidentes. Essa falta de uma correlação perfeita não é surpreendente, dado o que já suspeitávamos sobre o DL: ele é afetado não apenas pelo fato do quão provável é um evento de recombinação (isto é, onde ficam os *hot spots*), mas também pela idade da população, a frequência dos haplótipos originalmente presentes nos membros fundadores daquela população e se houve seleção positiva ou negativa para determinados haplótipos.

## MAPEAMENTO DE GENES DE DOENÇAS HUMANAS

### Por Que Mapear Genes de Doença?

Na medicina clínica, um estado de doença é definido por uma coleção de achados fenotípicos observados em um paciente ou grupo de pacientes. Designar essa doença como "genética" – e assim inferir a existência de um gene responsável ou que contribui para a doença – advém da análise genética detalhada, aplicando-se os princípios constantes dos Capítulos 7 e 8. No entanto, supor a existência de um gene ou genes dessa maneira *não* nos diz qual dos talvez 40.000 a 50.000 genes codificantes e não codificantes no genoma está envolvido, qual a função daquele gene ou genes, ou como aquele gene ou genes causam ou contribuem para a doença.

O mapeamento genético da doença é frequentemente um primeiro passo importante na identificação do gene ou

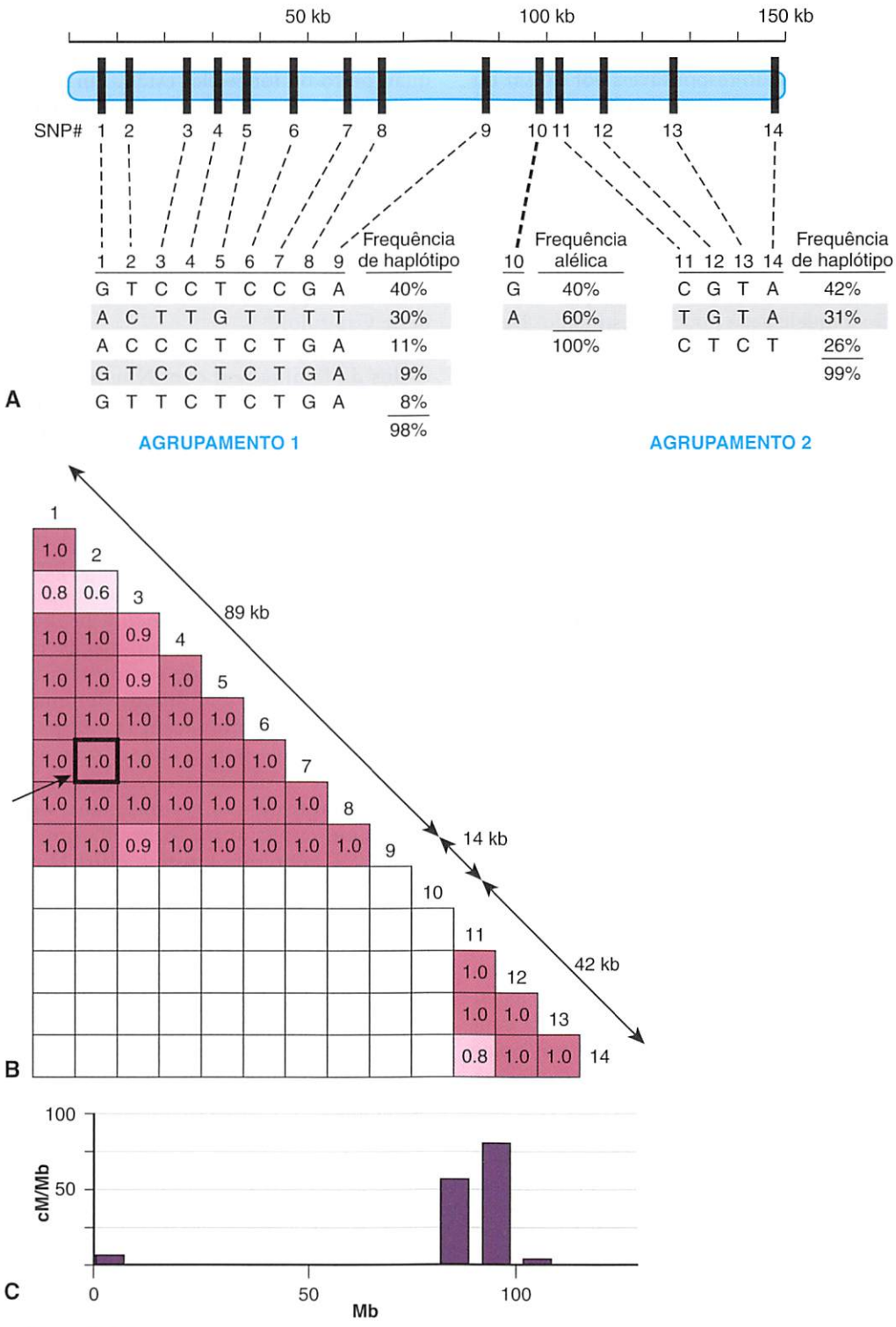



**Figura 10-9** A, Região de 145 kb do cromossomo 4 que contém 14 polimorfismos de nucleotídeo único (SNPs). No agrupamento 1, que contém os SNPs de 1 a 9, cinco dos $2^9 = 512$ haplótipos teoricamente possíveis são responsáveis por 98% de todos os haplótipos na população, refletindo desequilíbrio de ligação substancial (DL) entre esses *loci* de SNPs. Da mesma maneira, no agrupamento 2, apenas três dos $2^4 = 16$ haplótipos teoricamente possíveis que envolvem os SNPs de 11 a 14 representam 99% de todos os haplótipos encontrados. Em contrapartida, os alelos no SNP 10 encontram-se em equilíbrio de ligação com os SNPs no agrupamento 1 e agrupamento 2. B, Diagrama esquemático em que cada *quadro vermelho* contém a medição de pares do grau de DL entre dois SNPs (p. ex., a *seta* aponta para o quadro, *esboçado em preto*, contendo o valor de D' para os SNPs 2 e 7). Quanto maior o grau de DL, mais escura a cor do quadro, com valores máximos D' de 1,0 ocorrendo quando há DL completo. Dois blocos de DL são detectáveis, o primeiro contendo os SNPs de 1 a 9 e o segundo os SNPs entre 11 a 14. Entre os blocos, a região de 14 kb que contém o SNP 10 não mostra DL com os SNPs 9 e 11 circunvizinhos ou com qualquer um dos outros *loci* de SNPs. C, Gráfico da relação de distância de mapa e distância física (cM/Mb), mostrando que uma recombinação *hot spot* está presente na região entre o SNP 10 e o agrupamento 2, com valores de recombinação de cinquenta a sessenta vezes acima da média de, aproximadamente 1,15 cM/Mb para o genoma. *Veja Fontes & Agradecimentos.*

genes, nos quais variantes são responsáveis por causar ou aumentar a suscetibilidade à doença. O mapeamento do gene concentra a atenção sobre uma região do genoma, na qual se realiza uma análise sistemática de todos os genes da região para encontrar as mutações ou variantes que contribuem para a doença. Após a identificação do gene que abriga as variantes do DNA responsáveis por causar uma doença mendeliana ou aumentar a susceptibilidade a uma doença genética complexa, o espectro completo da variação naquele gene pode ser estudado. Desta maneira, podemos determinar o grau de heterogeneidade alélica, a penetrância de diferentes alelos, se existe uma correlação entre determinados alelos e vários aspectos do fenótipo (correlação genótipo-fenótipo) e a frequência de variantes causais de doenças ou predisponentes em várias populações.

Outros pacientes com o mesmo distúrbio ou outros semelhantes podem ser examinados para se observar se eles também abrigam ou não mutações no mesmo gene, o que indicaria que há heterogeneidade de *locus* para um determinado distúrbio. Após o gene e suas variantes genéticas naquele gene serem identificadas em indivíduos acometidos, métodos altamente específicos de diagnóstico, como diagnóstico pré-natal e triagem do portador, podem ser oferecidos aos pacientes e suas famílias.

As variantes associadas à doença podem ser, em seguida, modeladas em outros organismos, o que nos possibilita usar ferramentas genéticas, bioquímicas e fisiológicas poderosas para compreender melhor como a doença surge. Finalmente, armados com uma compreensão da função dos genes e como os alelos associados à doença afetam aquela função, podemos começar a desenvolver terapias específicas, como a terapia de reposição gênica, para evitar ou melhorar o distúrbio. Na verdade, grande parte do material nos próximos capítulos sobre a etiologia, patogenia, mecanismo e tratamento de várias doenças começa com o mapeamento genético. Aqui, examinamos as principais abordagens usadas para descobrir genes envolvidos na doença genética, tal como foi apresentado no início desse capítulo.

## Mapeamento dos Genes de Doenças Humanas por Análise de Ligação

### Determinação se Dois *Loci* estão Ligados

A análise de ligação é um método de mapeamento de genes que usa estudos de recombinação em famílias para determinar se dois genes apresentam ligação quando passados de uma geração para a seguinte. Usamos as informações do padrão de herança mendeliana conhecida ou suspeita (dominante, recessiva, ligada ao X) para determinar quais dos indivíduos na família herdaram um cromossomo recombinante ou não recombinante.

Para decidir se os dois *loci* estão ligados e, em caso afirmativo, quão perto ou distantes estão, contamos com duas informações. Primeiro, usando os dados da família em mãos, precisamos estimar θ, a frequência de recombinação entre os dois *loci*, porque isso vai nos dizer o quão perto ou longe eles estão. Em seguida, precisamos verificar se θ é estatisticamente significativamente diferente de 0,5, porque determinar se dois *loci* estão ligados é equivalente a perguntar se a fração de recombinação entre eles difere significativamente da fração esperada de 0,5 para *loci* não ligados. Estimar θ e, ao mesmo tempo, determinar a significância estatística de qualquer desvio de θ de 0,5, depende de uma ferramenta estatística chamada de **razão de probabilidade** (como discutido mais adiante neste Capítulo).

A análise de ligação começa com um conjunto de dados da família real com N indivíduos. Com base em um modelo de herança mendeliana, conte o número de cromossomos, r, que apresentam recombinação entre o alelo que causa a doença e alelos em vários *loci* polimórficos em torno do genoma (os chamados "marcadores"). O número de cromossomos que não apresentam uma recombinação é, portanto, N - r. A fração de recombinação θ pode ser considerada a probabilidade desconhecida, com cada meiose, que uma recombinação irá ocorrer entre os dois *loci*; a probabilidade de que não ocorra nenhuma recombinação é portanto 1 - θ. Pelo fato de cada meiose ser um evento independente, multiplica-se a probabilidade de uma recombinação, θ, ou de não recombinação, (1 - θ) para cada cromossomo. A fórmula para a chance (que é apenas a probabilidade) de se observar esse número de cromossomos recombinantes e não recombinantes quando θ é desconhecido é, portanto, fornecida por $\{N!/r!(N - r)!\}\theta^r (1 - \theta)^{(N - r)}$. (O termo fatorial, N!/r!(N - r)!, é necessário para explicar todas as possíveis ordens de nascimento em que as crianças recombinantes e não recombinantes podem aparecer no heredograma). Calcule uma segunda probabilidade baseada na hipótese nula de que os dois *loci* são não ligados, ou seja, fazer θ = 0,50. A razão entre a probabilidade de os dados da família que sustentam a ligação com θ desconhecido e a probabilidade de que os *loci* sejam não ligados é a probabilidade em favor de ligação e é fornecida por:

$$\frac{\text{Probabilidade dos dados se } loci \text{ forem ligados a uma distância } \theta}{\text{probabilidade dos dados se } loci \text{ forem não ligados } (\theta = 0,5)} = \frac{\{N!/r!(N-r)!\}\theta^r (1-\theta)^{(N-r)}}{\{N!/r!(N-r)!\}(\frac{1}{2})^r (\frac{1}{2})^{(N-r)}}$$

Felizmente, os termos fatoriais são sempre os mesmos no numerador e denominador da razão de verossimilhança e, portanto, eles se anulam mutuamente e podem ser ignorados. Se θ = 0,5, o numerador e o denominador são os mesmos e as probabilidades iguais a 1.

A teoria estatística nos diz que quando o valor da razão de verossimilhança para todos os valores de θ entre 0 e 0,5 é calculado, o valor de θ que dá o maior valor dessa razão de verossimilhança é, na verdade, a melhor estimativa da fração de recombinação que você pode fazer em vista dos dados e é referido como $\theta_{max}$. Por convenção, a razão de verossimilhança computada para diferentes valores de θ

em geral é expressa como $\log_{10}$ e é chamada de **escore do logaritmo da probabilidade (LOD *score*)** (Z) onde LOD é a abreviatura em inglês de "Logarithm of the ODds." O uso de logaritmos possibilita que as razões de verosimilhança calculadas de diferentes famílias sejam combinadas por adição simples em vez de ter de multiplicá-los juntos.

Como a análise do LOD *score* é realizada em famílias com distúrbios mendelianos? (Quadro) Retornemos à família mostrada na Figura 10-6, na qual a mãe tem uma forma autossômica dominante de **retinite pigmentosa**. Existem dezenas de diferentes formas dessa doença, sendo que muitas delas foram mapeadas em locais específicos dentro do genoma e para os genes para os quais foram agora identificados. Normalmente, quando uma nova família busca atendimento clínico, não se sabe que forma de RP o paciente tem. Nesta família, a mãe também é heterozigota para dois *loci* marcadores no cromossomo 7, o *locus* 1 em 7q distal e o *locus* 2 em 7p14. Suponha que sabemos (a partir de outros dados da família) que o alelo da doença *D* está em acoplamento com um alelo *A* no *locus* 1 e com o alelo *B* no *locus* 2. Dada essa fase, pode-se observar que houve recombinação entre RP e o *locus* 2 em apenas um dos seus oito filhos, sua filha II-3. Os alelos no *locus* da doença, no entanto, não apresentam qualquer tendência para seguir os alelos no *locus* 1 ou alelos em qualquer uma das outras centenas de loci marcadores testados nos outros autossomos. Assim, embora o *locus* de RP envolvido nesta família possa a principio ter sido mapeado em qualquer local do genoma humano, começa-se agora a suspeitar, com base nos dados da ligação, que o *locus* de RP responsável está na região do cromossomo 7, próximo do *locus* marcador 2.

Para fornecer uma avaliação quantitativa dessa suspeita, suponha que deixamos $\theta$ ser a fração de recombinação "verdadeira" entre RP e o *locus* 2, a fração que veríamos se tivéssemos números ilimitados de filhos para teste. A razão de verossimilhança para esta família é, portanto,

$$\frac{(\theta)^1(1-\theta)^7}{(\frac{1}{2})^1(\frac{1}{2})^7}$$

e alcança uma pontuação máxima de logaritmo de probabilidade de $Z_{max} = 1{,}1$ em $\theta_{max} = 0{,}125$.

O valor de $\theta$ que maximiza a razão de verossimilhança, $\theta_{max}$, pode ser a melhor estimativa que se pode fazer para determinados dados, mas qual o nível de qualidade dessa estimativa? A magnitude do LOD *score* fornece uma avaliação da qualidade de uma estimativa de $\theta_{max}$ que você fez. *Por convenção, um LOD score de +3 ou superior (equivalente a probabilidade superior a 1000:1 a favor da ligação) é considerado uma evidência firme de que dois loci estão ligados - ou seja, $\theta_{max}$ é estatisticamente muito diferente de 0,5.* Em nosso exemplo de RP, 7/8 da prole são não recombinantes e 1/8 são recombinantes. O $\theta_{max} = 0{,}125$, mas o LOD *score* é de apenas 1,1, o suficiente para levantar uma suspeita de ligação, mas insuficiente para comprovar ligação, porque $Z_{max}$ está muito aquém de 3.

### ANÁLISE DE LIGAÇÃO DE DOENÇAS MENDELIANAS

A análise de ligação é usada quando existe um modo particular de herança (autossômica dominante, autossômica recessiva ou ligada ao X) que explica o padrão de herança.

A análise do LOD *score* possibilita mapear os genes, nos quais mutações causam doenças que seguem herança mendeliana.

O LOD *score* fornece:

- A melhor estimativa da frequência de recombinação, $\theta_{max}$, entre um *locus* marcador e o *locus* da doença; e
- Uma avaliação de quão forte é a evidência de ligação naquele valor de $\theta_{máx}$. Valores de LOD *score* Z acima de 3 são considerados forte evidência.

A ligação em um $\theta_{max}$ específico entre um *locus* do gene para a doença e um marcador com local físico conhecido implica que o *locus* do gene da doença deve estar perto do marcador. Quanto menor o $\theta_{max}$, mais próximo o *locus* da doença está do *locus* marcador ligado.

## Combinação de Informações sobre o LOD *Score* nas Famílias

Da mesma maneira que cada meiose em uma família que produz uma prole não recombinante ou recombinante é um evento independente, assim também são as meioses que ocorrem em famílias diferentes. Podemos, portanto, multiplicar as probabilidades nos numeradores e denominadores de *odds ratio* de probabilidade de cada família juntos. Suponha que duas famílias adicionais com RP foram estudadas e uma não apresentou recombinação entre o *locus* 2 e RP em quatro filhos e outro não apresentou recombinação em cinco crianças. Os LOD *scores* individuais podem ser gerados para cada família e adicionados juntos (Tabela 10-1). Pelo fato de o LOD *score* máximo $Z_{max}$ exceder 3 a $\theta_{max} = \approx 0{,}06$, o gene de RP neste grupo de famílias está ligado ao *locus* 2 a uma distância de recombinação de $\approx 0{,}06$. Pelo fato de a localização genômica do *locus* marcador 2 ser conhecida por estar em 7p14, a RP nesta família pode ser mapeada na região 7p14 e provavelmente envolve o gene *RP9*, um dos *loci* já identificados para uma forma de RP autossômica dominante.

Se, no entanto, algumas das famílias que estão sendo usadas no estudo vão ter RP, devido a mutações em um *locus* diferente, os LOD *scores* entre as famílias vai divergir, com algumas apresentando uma tendência a ser positiva com valores pequenos de $\theta$ e outras apresentando LOD *scores* fortemente negativos nesses valores. Assim, na análise de ligação que envolve mais de uma família, uma heterogeneidade de *locus* insuspeita pode obscurecer o que pode ser uma evidência real de ligação em um subgrupo de famílias.

## Heredogramas de Fase Conhecida e Fase não Conhecida

No exemplo de RP que acabamos de discutir, supomos que conhecíamos a fase de alelos marcadores no cromossomo 7 na mãe acometida naquela família. Vamos agora ver as implicações de se conhecer a fase em mais detalhes.

**TABELA 10-1** LOD *Score* para Três Famílias com Retinite Pigmentosa

| | 0,00 | 0,01 | 0,05 | 0,06 | 0,07 | 0,10 | 0,125 | 0,20 | 0,30 | 0,40 |
|---|---|---|---|---|---|---|---|---|---|---|
| Família 1 | - | 0,38 | 0,95 | 1,00 | 1,03 | 1,09 | **1,1** | 1,03 | 0,80 | 0,46 |
| Família 2 | 1,2 | 1,19 | 1,11 | 1,10 | 1,08 | 1,02 | 0,97 | 0,82 | 0,58 | 0,32 |
| Família 3 | 1,5 | 1,48 | 1,39 | 1,37 | 1,35 | 1,28 | 1,22 | 1,02 | 0,73 | 0,39 |
| Total | - | 3,05 | 3,45 | 3,47 | 3,46 | 3,39 | 3,29 | 2,87 | 2,11 | 1,17 |

Indivíduo $Z_{max}$ para cada família é mostrado em negrito. O $Z_{máx}$ global = 3,47 a $\theta_{max}$ = 0,06.

Considere a família de três gerações com **neurofibromatose tipo 1 (NF1)** autossômica dominante (Caso 34) na Figura 10-10. A mãe acometida, II-2, é heterozigota tanto no *locus* NF1 (*D/d*) como no *locus* marcador (*A/a*), mas (como mostrado na Fig. 10-10A) não temos informações sobre o genótipo dos pais. As duas crianças acometidas receberam o alelo A juntamente com o alelo D da doença e a criança não acometida recebeu o alelo *a* juntamente com o alelo *d* normal. Sem conhecer a fase desses alelos na mãe, todos os três filhos são recombinantes ou todos os três são não recombinantes. Pelo fato de ambas as possibilidades serem igualmente prováveis na ausência de qualquer outra informação, nós consideramos a fase em seus dois cromossomos como sendo *D-a* e *d-A* metade das vezes *D-A* e *d-a* para a outra metade (que supõe que os alelos nesses haplótipos estão em equilíbrio de ligação). Para calcular a probabilidade global desse heredograma, em seguida adicionamos a probabilidade calculada supondo uma fase na mãe à probabilidade calculada supondo a outra fase. Portanto, a probabilidade global $=1/2\theta^0 (1-\theta)^3 + 1/2(\theta^3)(1-\theta)^0$ e a razão de verossimilhança para esse heredograma, em seguida, é:

$$\frac{\frac{1}{2}(1-\theta)^3(\theta)^0 + \frac{1}{2}(\theta^3)(1-\theta)^0}{\frac{1}{8}}$$

Conferindo um LOD *score* máximo de $Z_{max}$ = 0,602 a $\theta_{max}$ = 0.

Se, no entanto, a informação adicional do genótipo no avô materno I-1 torna-se disponível (como na Fig. 10-10B), a fase pode agora ser determinada como sendo *D-A* (ou seja, alelo D de NF1 estava no acoplamento com A no indivíduo II-2). Em função dessa nova informação, os três filhos agora podem ser definitivamente classificados como não recombinantes e já não temos mais de considerar a possibilidade de fase oposta. O numerador da razão de verossimilhança torna-se agora $(1-\theta)^3(\theta^0)$ e o logaritmo de probabilidade máximo $Z_{max}$ = 0,903 em $\theta_{max}$ = 0. Assim, o conhecimento da fase aumenta o poder dos dados disponíveis de testar a ligação.

## Mapeamento de Genes de Doenças Humanas por Associação

### Desenho de um Estudo de Associação

Uma abordagem completamente diferente para identificação da contribuição genética para a doença reside em encontrar *determinados alelos* que estejam associados à doença em uma amostra da população. Em contraste com a análise de ligação, essa abordagem não depende da existência de um padrão de herança mendeliana e é, portanto, mais adequado para descobrir as contribuições genéticas aos distúrbios com herança complexa (Cap. 8). A presença de um alelo em particular em um *locus* em frequência aumentada ou reduzida em indivíduos acometidos em comparação com controles é conhecida como uma **associação com a doença**. Há dois desenhos de estudos comumente usados para estudos de associação:

- **Estudos caso-controle.** Os indivíduos *com* a doença são selecionados em uma população, um grupo correspondente de controles *sem* doença é então selecionado e os genótipos de indivíduos nos dois grupos são determinados e utilizados para preencher uma tabela dois-por-dois (ver adiante).
- **Estudos de corte transversal ou de ou coorte.** Uma amostra aleatória de toda a população é escolhida

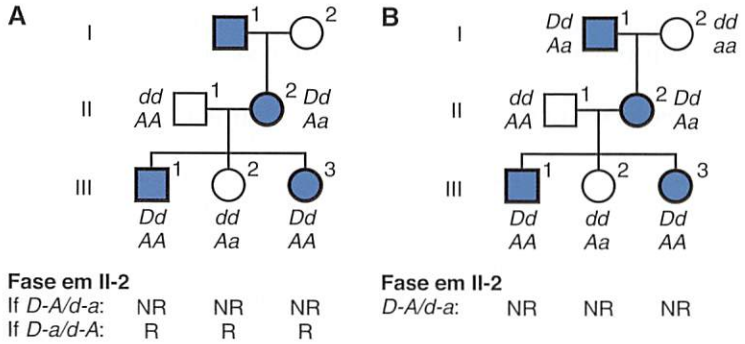



**Figura 10-10** Dois heredogramas de neurofibromatose autossômica dominante, tipo 1 (NF1). A, A fase do alelo da doença *D* e os alelos marcadores *A* e *a* no indivíduo II-2 é desconhecida. B, A disponibilidade de informações sobre o genótipo para a geração I possibilita uma determinação de que o alelo *D* da doença e alelo marcador *A* estão em acoplamento no indivíduo II-2. NR, não recombinante; R, recombinante.

e, em seguida, analisada para verificar se têm (corte transversal) ou, após serem acompanhados durante o tempo, desenvolvem (coorte) uma determinada doença; os genótipos de toda a população do estudo são determinados. O número de indivíduos com e sem doença e com e sem um alelo (ou genótipo ou haplótipo) de interesse é utilizado para preencher as células de uma tabela dois-por-dois.

## *Odds Ratio* e Riscos Relativos

Os dois tipos diferentes de estudos de associação relatam a força da associação, utilizando *odds ratio* ou risco relativo.

Em um **estudo caso-controle**, a frequência de um *determinado alelo ou haplótipo* (p. ex., para um haplótipo de antígeno leucocitário humano [HLA] ou um determinado alelo SNP ou haplótipo SNP) é comparada entre os indivíduos acometidos e não acometidos selecionados e uma associação entre a doença e o genótipo é calculada por uma ***odds ratio*** (OR) ou razão das chances.

| | Pacientes | Controles | Totais |
|---|---|---|---|
| Com marcador genético* | a | b | a + b |
| Sem marcador genético | c | d | c + d |
| Totais | a + c | b + d | |

*Um marcador genético pode ser um alelo, um genótipo ou um haplótipo.

Utilizando a tabela dois-por-dois, as chances de um portador do alelo desenvolver a doença é a relação (a/b) do número de portadores do alelo que desenvolvem a doença (a) e o número de portadores do alelo que não desenvolvem a doença (b). Da mesma maneira, as chances de um não portador desenvolver a doença é a razão (c/d) de não portadores que desenvolvem a doença (c) dividida pelo número de não portadores que não desenvolvem a doença (d). A *odds ratio* da doença é então a razão dessas probabilidades.

$$OR = \frac{\frac{a}{b}}{\frac{c}{b}} = \frac{ad}{bc}$$

Uma OR que difere de 1 significa que há uma associação do risco de doença com o marcador genético, enquanto OR = 1 significa que não há associação.

Alternativamente, se o estudo de associação foi concebido como um estudo **de corte transversal ou coorte**, a força de uma associação pode ser medida pelo **risco relativo (RR)**. O RR é a razão entre a proporção das pessoas com a doença que são portadoras de um alelo em particular ([a/(a +b)]) e a proporção daqueles sem a doença que são portadores daquele alelo ([c/(c + d)]).

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Novamente, um RR que difere de 1 significa que há uma associação de risco de doença com o marcador genético, enquanto RR = 1 significa que não há associação. (O risco relativo RR introduzido aqui não deve ser confundido com $\lambda_p$, a **relação de risco em parentes**, que foi discutido no Capítulo 8. $\lambda_r$ é a prevalência de um fenótipo de doença em particular em parentes de um indivíduo acometido *versus* aquele na população geral.)

Para doenças raras (ou seja, a <b e c <d), um desenho caso-controle com cálculo da OR é o melhor, porque qualquer amostra aleatória de uma população provavelmente não contém números suficientes de indivíduos acometidos para ser adequado para um desenho de estudo de corte transversal ou de coorte. Observe, entretanto, que quando uma doença é rara e o cálculo de uma OR em um estudo de caso controle é a única abordagem prática, a OR é uma boa aproximação para um RR. (Examine a fórmula para RR e se convença de que, quando a <b e c <d (a + b) ≈ b e (c + d) ≈ d, e, portanto, RR ≈ OR.)

A informação obtida em um estudo de associação vem em duas partes. A primeira é a **magnitude da associação** em si: quanto mais o RR ou OR divergirem de 1, maior é o efeito da variação genética na associação. No entanto, uma OR ou um RR para uma associação é uma medida estatística e requer um teste de significância estatística. A **significância** de qualquer associação pode ser avaliada simplesmente perguntando com um teste de qui-quadrado, se as frequências alélicas (a, b, c e d, na tabela dois a dois) diferem significativamente do que seria esperado se não houvesse nenhuma associação (ou seja, se o OR ou RR fossem iguais a 1,0). Uma maneira comum de expressar se há significância estatística para uma estimativa de OR ou RR é promover um **intervalo de confiança** de 95% (ou 99%). O intervalo de confiança é o intervalo dentro do qual seria de se esperar que a OR ou o RR caísse de 95% (ou 99%) do tempo ao acaso isoladamente em uma amostra colhida da população. Se um intervalo de confiança exclui o valor 1,0, então a OR ou RR desvia significativamente do que seria esperado se não houvesse nenhuma associação com o *locus* marcador que está sendo testado, e a hipótese nula de não associação pode ser rejeitada em nível correspondente de significância. (Mais adiante neste capítulo, vamos explicar por que um nível de 0,05 ou 0,01 é inadequado para a avaliação da significância estatística quando *múltiplos loci* marcadores no genoma são simultaneamente testados para a associação.)

Para ilustrar essas abordagens, primeiro consideramos um estudo caso controle de **trombose venosa cerebral** (TVC), que nós introduzimos no Capítulo 8. Neste estudo, suponha que um grupo de 120 pacientes com TVC e 120 controles semelhantes foram genotipados para o alelo 20210G > A no gene da protrombina (Cap. 8).

| | Pacientes com TVC | Controles sem TVC | Totais |
|---|---|---|---|
| Alelo 20210G > A presente | 23 | 4 | 27 |
| Alelo 20210G > A ausente | 97 | 116 | 213 |
| Total | 120 | 120 | 240 |

TVC, trombose venosa cerebral.

Pelo fato de este ser um estudo caso-controle, iremos calcular uma *odds ratio*: OR = (23/4)/(97/116) = ≈6,9 com limites de 95% de confiança de 2,3 a 20,6. Há claramente um tamanho de efeito substancial de 6,9 e limites de 95% de confiança que excluem 1,0, demonstrando deste modo que existe uma associação forte e estatisticamente significativa entre o alelo 20210G > A e a TVC. Simplificando, indivíduos portadores do alelo de protrombina 20210G > A têm probabilidades quase sete vezes maiores de ter a doença do que aqueles que não são portadores desse alelo.

Para ilustrar um estudo de coorte longitudinal em que a RR, em vez de uma OR, pode ser calculado, considere **a miopatia induzida por estatinas**, uma reação medicamentosa adversa rara, mas bem reconhecida, que pode se desenvolver em alguns indivíduos durante a terapia com estatinas para baixar o colesterol. Em um estudo, os indivíduos incluídos no estudo de proteção cardíaca foram randomizados para receber 40 mg do fármaco de estatina sinvastatina ou placebo. Mais de 16.600 participantes expostos à estatina foram genotipados para uma variante (Val174Ala) no gene *SLCO1B1*, que codifica um transportador hepático do fármaco e foram observados para o desenvolvimento de resposta adversa ao medicamento. Do grupo completo genotipado exposto à estatina, 21 desenvolveram miopatia. O exame dos seus genótipos mostrou que o RR para o desenvolvimento de miopatia associada à presença do alelo Val174Ala é de cerca de 2,6, com limites de confiança de 95% de 1,3 a 5,1. Assim, aqui há uma *associação estatisticamente significativa entre o alelo Val174Ala e a miopatia induzida por estatinas*; aqueles portadores deste alelo apresentam risco moderadamente aumentado de desenvolver esta reação medicamentosa adversa em relação àqueles que não são portadores desse alelo.

Um equívoco comum referente a um estudo de associação é que quanto mais significativo o valor de *P*, mais forte a associação. Na verdade, um valor de *P* significativo para uma associação *não* fornece informações relativas à magnitude do efeito de um alelo associado na suscetibilidade à doença. A significância é uma medida estatística que descreve quão provável é que a amostra de população utilizada para o estudo de associação poderia ter produzido uma OR ou RR observada que difere de 1,0 simplesmente por acaso apenas. Em contraste, a magnitude real de OR ou RR – *quanto* diverge de 1,0 – é uma medida do impacto que uma variante em particular (ou genótipo ou haplótipo) tem sobre o aumento ou redução da suscetibilidade da doença.

## Estudos de Associação Genômica Ampla

### O Mapa de Haplótico (HapMap)

Por muitos anos, os estudos de associação para genes de doenças humanas eram limitados a conjuntos particulares de variantes em conjuntos restritos de genes escolhidos seja por conveniência ou porque foram considerados envolvidos em uma via fisiopatológica relevante para uma doença e, assim, pareciam ser **genes candidatos** lógicos para a doença sob investigação. Assim, muitos desses estudos de associação foram realizados antes da era do Projeto Genoma Humano com o uso dos *loci* de HLA ou do grupo sanguíneo, por exemplo, porque esses *loci* eram altamente polimórficos e facilmente genotipados em estudos caso-controle. Idealmente, contudo, seria interessante ser capaz de testar sistematicamente uma associação entre qualquer doença de interesse e *cada* um das dezenas de milhões de alelos raros e comuns no genoma de uma maneira imparcial, sem qualquer pré-conceito de quais genes e variantes genéticas poderiam estar contribuindo para a doença.

As análises de associação em uma escala genômica são chamadas de estudos de associação genômica ampla (do inglês, *genome-wide association studies*), conhecidos por seu acrônimo **GWAS**. Tal empreendimento para *todas* as variantes conhecidas é impraticável por muitas razões, mas pode ser aproximado pela genotipagem de casos e controles para meras 300.000 a 1 milhão de variantes isoladas localizadas em todo o genoma para procurar associação com a doença ou a característica em questão. O sucesso dessa abordagem depende da exploração do DL porque desde que uma variante responsável por alterar a suscetibilidade a doenças esteja em DL com uma ou mais das variantes genotipadas dentro de um bloco de DL, uma associação positiva deve ser detectável entre aquela doença e os alelos no bloco de DL.

O desenvolvimento desse conjunto de marcadores levou ao lançamento do **Projeto de Mapeamento de Haplótipos (HapMap)**, um dos maiores esforços genômicos humanos para acompanhar a conclusão do Projeto Genoma Humano. O Projeto HapMap começou em quatro grupos geograficamente distintos – uma população principalmente europeia, uma população do Oeste Africano, uma população chinesa Han e uma população do Japão - e incluiu a coleta e caracterização de milhões de *loci* de SNPs e métodos de desenvolvimento para genotipá-los rapidamente e de maneira barata. Desde aquela época, o sequenciamento de genoma completo foi aplicado a muitas populações no chamado **Projeto 1000 Genomas**, resultando em uma enorme expansão da base de dados de variantes de DNA disponíveis para GWAS, com diferentes populações em todo o globo.

## Mapeamento Gênico por Estudos de Associação Genômica Ampla em Traços Complexos

O objetivo do HapMap não era apenas o de reunir informações básicas sobre a distribuição de DL em todo o genoma humano. Seu objetivo principal era fornecer uma ferramenta poderosa nova para encontrar as variantes genéticas que contribuem para as doenças humanas e outros traços, tornando possível uma aproximação com uma associação genômica ampla idealizada em grande escala. O princípio impulsionador por trás dessa abordagem é simples: a *detecção de uma associação com alelos dentro de um bloco de DL aponta a região genômica dentro do bloco como propensa a conter o alelo associado à doença*. Consequentemente,

embora a abordagem tipicamente não indique a variante *real* funcionalmente responsável pela associação com a doença, esta região será o local a concentrar estudos adicionais para encontrar a variante alélica que *está* funcionalmente envolvida no processo de doença em si.

Historicamente, a análise detalhada de condições associadas a variantes de alta densidade nas regiões de HLA de classe I e classe II (Fig 8-10) exemplificaram essa abordagem (Quadro). No entanto, com as dezenas de milhões de variantes atualmente disponíveis em diferentes populações, esta abordagem pode ser ampliada para examinar a base genética de praticamente *qualquer* doença ou característica complexa. Na verdade, até o momento, milhares de GWAS descobriram um número enorme de variantes de ocorrência natural associadas a uma variedade de doenças multifatoriais geneticamente complexas, que variam de diabetes e doença intestinal inflamatória até artrite reumatoide e câncer, bem como para os traços como estatura e pigmentação. Uma pesquisa para descobrir a base biológica subjacente para essas associações estará em curso nos próximos anos.

### ANTÍGENO LEUCOCITÁRIO HUMANO E ASSOCIAÇÃO COM DOENÇAS

Dentre as mais de mil associações genômicas com traço ou doença a partir de todo o genoma, a região com a maior concentração de associações a diferentes fenótipos é a região do antígeno leucocitário humano (HLA). Além da associação de alelos e haplótipos específicos ao **diabetes tipo 1**, discutida no Capítulo 8, a associação de vários polimorfismos de HLA foi demonstrada para uma ampla gama de condições, sendo que a maioria, mas não todas elas, é **autoimune**, ou seja, está associada a uma resposta imune anormal aparentemente dirigida contra um ou mais autoantígenos. Essas associações são consideradas relacionadas com a variação na resposta imune resultante de polimorfismo em genes da resposta imune.

A base funcional da maioria das associações HLA-doença é desconhecida. Moléculas HLA são parte integrante do reconhecimento de células T de antígenos. Acredita-se que diferentes alelos HLA polimórficos resultam em variação estrutural nessas moléculas da superfície celular, conduzindo a diferenças na capacidade das proteínas de interagir com o antígeno e o receptor de célula T na iniciação de uma resposta imune, afetando assim esses processos críticos como a imunidade contra infecções e a autotolerância para evitar a autoimunidade.

A **espondilite anquilosante**, doença inflamatória crônica da coluna vertebral e das articulações sacroilíacas, é um exemplo. Mais de 95% das pessoas com espondilite anquilosante são positivas para HLA-B27; o risco de desenvolver espondilite anquilosante é 150 vezes maior para as pessoas que têm alelos HLA-B27 do que para aquelas que não têm. Esses alelos levam a erros no dobramento da cadeia pesada de HLA-B27 e apresentação ineficiente de antígenos.

Em outros distúrbios, a associação entre um alelo HLA específico ou haplótipo e uma doença não é causada por diferenças funcionais nos genes da resposta imune em si. Em vez disso, a associação é causada por um determinado alelo presente em uma frequência muito alta em cromossomos que também têm mutações causais de doenças em outro gene na principal região complexa de histocompatibilidade. Um exemplo é a **hemocromatose**, um distúrbio comum de sobrecarga de ferro. Mais de 80% dos pacientes com hemocromatose são homozigotos para uma mutação comum, a Cys282Tyr, no gene da hemocromatose (*HFE*) e possuem alelos HLA-A*0301 em seu *locus* de *HLA-A*. A associação, contudo, não é resultado de HLA-A*0301. A *HFE* está envolvida no transporte ou metabolismo de ferro no intestino; *HLA-A*, como um gene de resposta imune de classe I, não têm efeito sobre o transporte de ferro. A associação ocorre devido à proximidade dos dois *loci* e DL entre uma mutação de HFE Cys282Tyr e o alelo *A*0301* em *HLA-A*.

## Armadilhas no Desenho e na Análise de GWAS

Métodos de associação são ferramentas poderosas para identificar com precisão os genes que contribuem para a doença genética por meio da demonstração não só dos genes, mas também dos alelos específicos responsáveis. Eles são também relativamente fáceis de realizar porque são necessárias apenas amostras de um conjunto de indivíduos acometidos não aparentados e controles, e não é necessário realizar estudos familiares trabalhosos nem coleta de amostras de vários membros de um heredograma.

Os estudos de associação devem ser interpretados com cautela, contudo. Uma grave limitação dos estudos de associação é o problema da associação totalmente artefactual causada pela **estratificação da população** (Cap. 9). Se uma população é estratificada em subpopulações separadas (p. ex., por etnia ou religião) e membros de uma subpopulação raramente se relacionam com membros de outras subpopulações, então uma doença que seja mais comum em uma subpopulação, seja por que razão for, pode parecer (incorretamente) estar associada a quaisquer alelos que também venham a ser mais comuns naquela subpopulação do que na população como um todo. A associação factícia decorrente da estratificação da população pode ser minimizada, no entanto, pela seleção cuidadosa dos controles pareados. Em particular, uma forma de controle de qualidade é certificar-se de que os casos e os controles têm frequências semelhantes de alelos, cujas frequências são conhecidas por diferir acentuadamente entre as populações (**marcadores informativos de ancestralidade**, como discutido no Cap. 9). Se as frequências observadas em casos e controles forem semelhantes, então a estratificação insuspeita ou oculta é improvável.

Além do problema de a estratificação produzir associações falso positivas, os resultados falsos positivos no GWAS podem também surgir se um teste inadequadamente vago para significância estatística for aplicado. Isso ocorre porque, como o número de alelos que está sendo testado para uma associação a doença aumenta, a chance de encontrar associações *ao acaso isoladamente* também aumenta, um conceito em estatística conhecido como o problema do **teste de hipóteses múltiplas**. Para compreender por que o ponto de corte para significância estatística deve ser muito mais rigoroso quando múltiplas hipóteses estão sendo testadas, imagine jogar uma moeda 50 vezes e ela cair em cara 40 vezes. Tal resultado altamente incomum tem uma probabilidade de ocorrência de apenas uma vez em

aproximadamente 100.000 vezes. No entanto, se o mesmo experimento for repetido milhões de vezes, as chances são superiores a 99.999% de que *pelo menos* uma tentativa de jogar a moeda, dentre os milhões de tentativas realizadas, resulte em 40 ou mais caras!

Assim, mesmo eventos raros que ocorrem por acaso isoladamente em um experimento tornam-se frequentes quando o experimento é repetido muitas vezes. É por isso que quando se faz o teste para uma associação com centenas de milhares de milhões de variantes de todo o genoma, dezenas de milhares de variantes poderiam aparecer associadas a $P < 0{,}05$ *ao acaso isoladamente*, fazendo um corte típico para significância estatística de $P < 0{,}05$ demasiado baixo para apontar para uma verdadeira associação. Em vez disso, um nível de significância de $P < 5 \times 10^{-8}$ é considerado mais apropriado para o GWAS que testa centenas de milhares a milhões de variantes. Mesmo com pontos de corte apropriadamente rigorosos para significância em todo o genoma, contudo, resultados falso positivos, devido apenas ao acaso, ainda ocorrerão. Para considerar isso, um GWAS devidamente realizado em geral inclui um **estudo de replicação** em um grupo diferente, completamente independente de indivíduos para mostrar que alelos próximos do mesmo *locus* estão associados. Uma ressalva, no entanto, é que alelos que apresentam associação podem ser diferentes em diferentes grupos étnicos.

Finalmente, é importante ressaltar que, se for encontrada uma associação entre uma doença e um alelo marcador polimórfico que é parte de um mapa de haplótipos denso, *não se pode* inferir que há um papel funcional para esse alelo marcador no aumento da suscetibilidade à doença. Devido à natureza do DL, *todos* os alelos em DL com um alelo em um *locus* envolvido na doença apresentará uma associação aparentemente positiva, independentemente de terem qualquer relevância funcional na predisposição à doença. Uma associação baseada em DL ainda é bastante útil, no entanto, porque para os alelos de marcadores polimórficos parecerem associados, os alelos de marcadores polimórficos associados provavelmente estarão localizados no bloco de DL que também abriga o *locus* real da doença.

A comparação das características, pontos fortes e pontos fracos de métodos de ligação e associação para mapeamento do gene da doença estão resumidos no Quadro.

## DO MAPEAMENTO GÊNICO À IDENTIFICAÇÃO DO GENE

A aplicação do mapeamento gênico à genética médica usando abordagens descritas na seção anterior alcançou muitos sucessos espetaculares. Essa estratégia levou à identificação dos genes associados a milhares de distúrbios mendelianos e um número crescente de genes e alelos associados a distúrbios geneticamente complexos. O poder dessas abordagens aumentou grandemente com a introdução de tecnologias altamente eficientes e menos caras para a análise do genoma.

### COMPARAÇÃO DE LIGAÇÃO E MÉTODOS DE ASSOCIAÇÃO

| Ligação | Associação |
|---|---|
| • Segue herança de um traço de doença e regiões do genoma de indivíduo para indivíduo em heredogramas familiares<br>• Procura regiões do genoma que abrigam alelos da doença; usa variantes polimórficas apenas como uma maneira de marcar qual região um indivíduo herdou de qual progenitor<br>• Utiliza centenas a milhares de marcadores polimórficos em todo o genoma<br>• Não é projetado para encontrar a variante específica responsável ou predisponente à doença; só pode demarcar onde a variante pode ser encontrada (geralmente) dentro de uma ou algumas megabases<br>• Depende dos eventos de recombinação que ocorrem nas famílias durante apenas algumas gerações para possibilitar a medição da distância genética entre um gene da doença e marcadores polimórficos nos cromossomos<br>• Requer amostragem de famílias, não apenas pessoas acometidas pela doença<br>• Perde potência quando a doença tem herança complexa com substancial falta de penetrância<br>• Mais frequentemente usado para mapear mutações causais de doenças com efeitos fortes o suficiente para causar um padrão de herança mendeliana | • Testes para alteração de frequência de alelos ou haplótipos específicos em indivíduos acometidos comparados com controles em uma população<br>• Examina alelos ou haplótipos específicos para sua contribuição para a doença<br>• Usa qualquer lugar de alguns marcadores em genes almejados a centenas de milhares de marcadores para análises de todo o genoma<br>• Pode ocasionalmente identificar a variante que é realmente funcionalmente responsável pela doença; mais frequentemente, define um haplótipo que contém a doença em um intervalo de 1 a 10 kb (em geral)<br>• Depende de se encontrar um conjunto de alelos, incluindo o gene da doença, que permaneceram juntos durante muitas gerações devido a uma *falta* de eventos de recombinação entre os marcadores<br>• Pode ser realizada em amostras caso-controle ou de coorte de populações<br>• É sensível ao artefato de estratificação da população, embora possa ser controlado por desenhos caso-controle adequados ou o uso de abordagens baseadas na família<br>• É a melhor abordagem para encontrar variantes com efeito pequeno que contribuem para traços complexos |

Nesta seção, descrevemos como métodos genéticos e genômicos levaram à identificação dos genes envolvidos em dois distúrbios, um usando primeiro a análise de ligação e DL para refinar a localização do gene responsável pela doença autossômica recessiva comum **fibrose cística** (FC) (Caso 12) e um usando GWAS para encontrar múltiplas variantes alélicas nos genes que aumentam a susceptibilidade à **degeneração macular relacionada com a idade** (DMI) (Caso 3), um distúrbio devastador que rouba a visão de adultos.

## Encontrar Gene em um Distúrbio Mendeliano Comum por Mapeamento de Ligação

### Exemplo: Fibrose Cística

Por causa de sua frequência relativamente alta, particularmente em populações brancas, e a quase completa falta de compreensão das anormalidades subjacentes à sua patogenia, a FC representou uma excelente candidata para a identificação do gene responsável pela utilização de ligação para encontrar a localização do gene, em vez de utilizar qualquer informação sobre o próprio processo de doença. As amostras de DNA de quase 50 famílias com FC foram analisadas para ligação entre a FC e centenas de marcadores de DNA em todo o genoma até a ligação de FC a marcadores no braço longo do cromossomo 7 ser finalmente identificada. A ligação a marcadores adicionais do DNA em 7q31-q32 estreitou a localização do gene da FC a uma região de aproximadamente 500 kb do cromossomo 7.

***Desequilíbrio de Ligação na Fibrose Cística.*** Neste ponto, no entanto, surgiu uma característica importante da genética da FC: embora os marcadores ligados mais próximos estivessem a alguma distância do gene da FC, tornou-se claro que havia DL significativo entre o *locus* da doença e um determinado haplótipo nos *loci* fortemente ligados à doença. As regiões com o maior grau de DL foram analisadas para sequências de genes, levando ao isolamento do gene da FC em 1989. Como descrito em detalhes no Capítulo 12, o gene responsável, que foi chamado de regulador de condutância transmembranar da fibrose cística (*CFTR*), apresentou um espectro interessante de mutações. Uma deleção de 3-pb (ΔF508) que removeu uma fenilalanina na posição 508 na proteína foi encontrada em cerca de 70% de todos os alelos mutantes de FC nas populações do norte da Europa, mas nunca entre alelos normais neste *locus*. Embora estudos posteriores tenham demonstrado muitas centenas de alelos *CFTR* mutantes em todo o mundo, foi a alta frequência da mutação de ΔF508 nas famílias usadas para mapear o gene da FC e do DL entre ele e os alelos nos *loci* de marcadores polimórficos próximos que comprovaram ser úteis na identificação final do gene *CFTR*.

O mapeamento do *locus* de FC e clonagem do gene *CFTR* possibilitou uma ampla gama de avanços da pesquisa e aplicações clínicas, de fisiopatologia básica a diagnóstico molecular para o aconselhamento genético, diagnóstico pré-natal, modelos animais, e finalmente tentativas contínuas atuais para tratar o distúrbio (Cap. 12).

## Encontrar os Genes que Contribuem para uma Doença Complexa por Associação Genômica Ampla

**Exemplo*: Degeneração Macular Relacionada com a Idade.*** A DMI é uma doença degenerativa progressiva da porção da retina responsável pela visão central. Ela causa cegueira em 1,75 milhões de americanos com mais de 50 anos de idade. A doença é caracterizada pela presença de drusas, que são depósitos extracelulares clinicamente visíveis, distintos de proteína e lipídeos atrás da retina na região da mácula (Caso 3). Embora haja ampla evidência de uma contribuição genética para a doença, a maioria dos indivíduos com DMI não está em famílias em que há um provável padrão mendeliano de herança. As contribuições ambientais também são importantes, como mostrado pelo aumento do risco de DMI em indivíduos tabagistas em comparação com não fumantes.

Os GWAS iniciais caso-controle da DMI revelaram uma associação de dois SNPs comuns próximos do gene do fator do complemento H (CFH). O haplótipo em risco mais frequente que contém esses alelos foi observado em 50% dos casos *versus* apenas 29% de controles (OR = 2,46; 95% intervalo de confiança [IC], 1,95-3,11). A homozigosidade para este haplótipo foi encontrada em 24,2% dos casos, em comparação com apenas 8,3% dos controles (OR = 3,51; IC 95%, 2,13-5,78). A pesquisa através dos SNPs dentro do bloco de DL que contém o haplótipo associado à DMI revelou um SNP não sinônimo no gene *CFH* que substituiu a tirosina por histidina na posição 402 da proteína *CFH* (Tyr402His). A alteração Tyr402His, que tem uma frequência alélica de 26% a 29%, em populações caucasianas e africanas, apresentou uma associação ainda mais forte com a DMI do que os dois SNPs que mostraram uma associação nos GWAS originais.

Como as drusas contêm fatores do complemento e o CFH é encontrado nos tecidos da retina ao redor das drusas, acredita-se que a variante Tyr402His é menos protetora contra a inflamação que é considerada responsável pela formação de drusas e danos na retina. Assim, a Tyr402His provavelmente é a variante no *locus* de *CFH* responsável por aumentar o risco para DMI.

Os GWAS mais recentes de DMI que utilizam mais de 7.600 casos e mais de 50.000 controles e milhões de variantes de todo o genoma revelaram que os alelos em um mínimo de 19 *loci* estão associados à DMI, com significância em todo o genoma de $P < 5 \times 10^{-8}$. Uma maneira popular de resumir GWAS na forma de gráfico é traçar os níveis de significância de $-\text{Log}_{10}$ para cada variante associada no que é chamado de "gráfico de Manhattan," porque se considera que possui uma semelhança um tanto fantasiosa com a linha do horizonte da cidade de Nova York (Fig. 10-11). As OR para DMI dessas variantes variam de um máximo de 2,76 para um gene de função desconhecida, *ARMS2*, e 2,48 para *CFH* a 1,1 para muitos outros genes envolvidos em múltiplas vias, incluindo o sistema complemento, aterosclerose, formação de vasos sanguíneos, e outros.

Neste exemplo de DMI, uma doença complexa, os GWAS levaram à identificação de SNPs comuns, fortemente

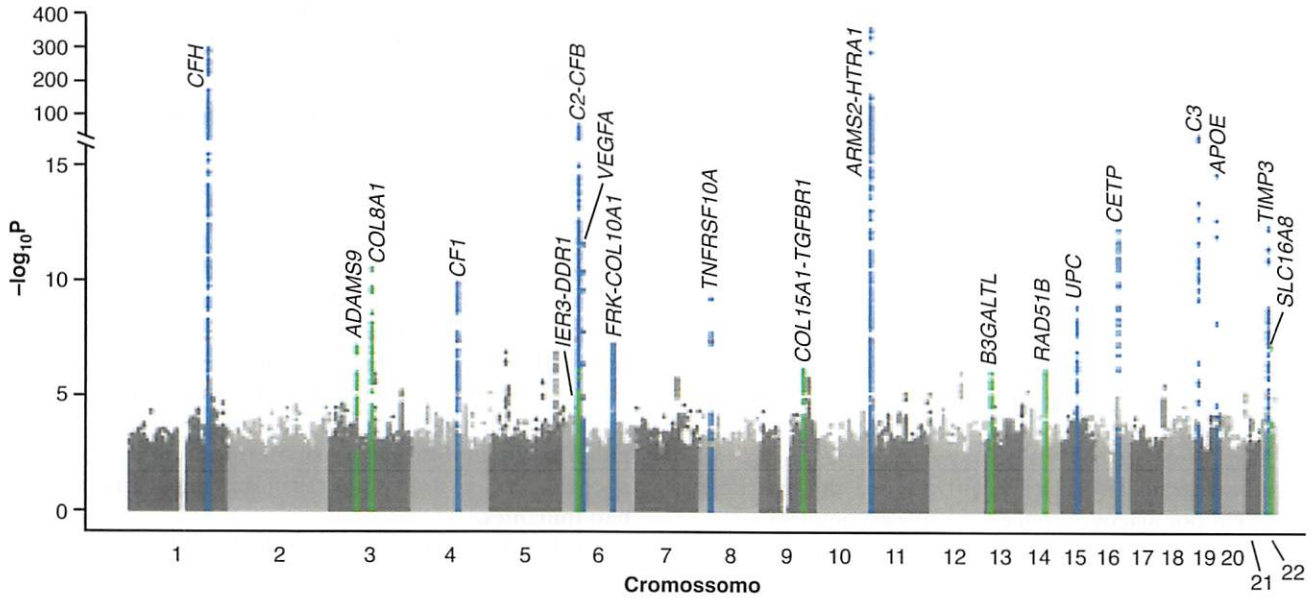



**Figura 10-11** **"Gráfico de Manhattan" de estudos de associação genômica ampla (GWAS) da degeneração macular relacionada com a idade usando aproximadamente 1 milhão de alelos de polimorfismos de nucleotídeo único (SNPs) do genoma inteiro localizados ao longo de todos os 22 autossomos no eixo-x.** Cada *ponto azul* representa a significância estatística (expressa como $-\log_{10}(P)$ colocada em gráfico no eixo y), confirmando uma associação anteriormente conhecida; *pontos verdes* são a significância estatística para novas associações. A descontinuidade no eixo y é necessária porque algumas das associações têm valores *P* extremamente pequenos $<1 \times 10^{-16}$. *Veja Fontes & Agradecimentos.*

associados, que por sua vez estavam em DL com um SNP codificante comum no gene que parece ser a variante funcional envolvida na doença. Esta descoberta, por sua vez, conduziu à identificação de outros SNPs na cascata do complemento e em outros lugares que podem também predispor a ou proteger contra a doença. Somados, esses resultados fornecem indícios importantes para a patogenia da DMI e sugerem que a via do complemento pode ser um alvo frutífero para novas terapias. Igualmente interessante é que o GWAS revelou que um novo gene de função desconhecida, o *ARMS2*, também está envolvido, abrindo, assim, uma linha inteiramente nova de pesquisa sobre a patogenia da DMI.

## Importância das Associações Descobertas com GWAS

Há uma discussão vigorosa com relação à interpretação dos resultados de GWAS e seu valor como uma ferramenta para estudos da genética humana. O debate surge principalmente a partir de um mal-entendido do que um OR ou RR significa. É verdade que muitos GWAS realizados de maneira apropriada produzem associações significativas, mas de tamanho de efeito muito modesto (semelhante ao OR de 1,1 mencionado há pouco para DMI). Na verdade, associações significativas de tamanho de efeito cada vez menores tornaram-se mais comuns à medida que tamanhos de amostras cada vez maiores são usadas, o que possibilita a detecção de associações no genoma inteiro estatisticamente significativas com OR ou RR cada vez menores. Isto levou à sugestão de que os GWAS são de pouco valor porque o tamanho do efeito da associação, medido pelo OR ou RR, é demasiadamente pequeno para o gene e a via implicados por essa variante para ser importante na patogenia da doença. Este raciocínio é falho por dois motivos.

Em primeiro lugar, as OR são uma medida do impacto de um alelo específico (p. ex., o alelo *CFH* Tyr402His para DMI) sobre as vias patogenéticas complexas, como a via do complemento alternativo, do qual o CFH é um componente. A sutileza desse impacto é determinada pela maneira como esse alelo perturba a função biológica do gene onde está localizado, e não pelo fato de o gene que abriga aquele alelo poder ou não ser importante na patogenia da doença. Em doenças autoimunes, por exemplo, estudos de pacientes com um número de doenças autoimunes diferentes, como **artrite reumatoide, lúpus eritematoso sistêmico** e **doença de Crohn**, revelam associações modestas, mas com algumas das mesmas variantes, sugerindo que existem vias comuns que levam a estas doenças distintas, mas relacionadas, que provavelmente serão bastante esclarecedoras em estudos de sua patogenia (Quadro).

Em segundo lugar, mesmo se o tamanho do efeito de qualquer variante for pequeno, os GWAS demonstram que muitos desses distúrbios são de fato extremamente poligênicos, ainda mais poligênicos do que suspeitava anteriormente, com milhares de variantes, sendo que a maior parte delas contribui apenas um pouco (OR entre 1,01 e 1,1) para a suscetibilidade à doença em si, mas, em conjunto, são responsáveis por uma fração substancial do agrupamento observado destas doenças em determinadas famílias (Cap. 8).

Embora a observação do tamanho de efeito modesto para a maioria dos alelos encontrados por GWAS esteja correta, ele negligencia um achado importante e talvez mais fundamental do GWAS: *a arquitetura genética de algumas das doenças complexas mais comuns estudadas até o momento pode envolver centenas de milhares de loci que abrigam variantes de efeito pequeno em muitos genes e vias*. Estes genes e vias são importantes para a nossa compreensão de como doenças complexas ocorrem, mesmo que cada alelo exerça apenas efeitos sutis sobre a regulação gênica ou função da proteína e tem apenas um efeito modesto na suscetibilidade à doença em uma base por alelo.

Assim o GWAS continua sendo uma ferramenta de pesquisa de genética humana importante para dissecar as muitas contribuições para doenças complexas, independente de as variantes individuais encontradas como associadas à doença aumentarem substancialmente o risco em indivíduos portadores desses alelos (Cap. 16). Esperamos que muitas outras variantes genéticas responsáveis pelas doenças complexas sejam identificadas de maneira bem-sucedida por associação genômica ampla e que o sequenciamento profundo das regiões que apresentam associações com doenças desvendem as variantes ou coleções de variantes funcionalmente responsáveis por associações de doenças. Tais achados devem nos fornecer uma compreensão poderosa e alvos terapêuticos potenciais para muitas das doenças comuns que causam tanta morbidade e mortalidade na população.

### DO GWAS AO PHEWAS

Em estudos de associação genômica ampla (GWAS), explora-se a base genética para um determinado fenótipo, doença ou traço procurando por associações com grandes coleções, sem viés, de marcadores de DNA do genoma inteiro. Mas pode-se fazer o inverso? Podem-se descobrir as potenciais *ligações fenotípicas* com variantes do genoma procurando associações com coleções grandes sem viés de fenótipos do "fenoma" inteiro? Até agora, os resultados dessa abordagem parecem ser altamente promissores.

Em uma abordagem apelidada de estudos de associação fenômica ampla (**PheWAS**), variantes genéticas são testadas para a associação, não apenas com um fenótipo especial de interesse (p. ex., artrite reumatoide ou pressão arterial sistólica acima de 160 mmHg), mas com todos os fenótipos clinicamente relevantes e valores laboratoriais encontrados em **prontuários eletrônicos** (**PE**). Desta maneira, é possível buscar associações novas e imprevistas sem vieses, utilizando algoritmos de pesquisa, códigos de faturamento, mineração de texto aberto para consulta de todas as entradas eletrônicas, que estão rapidamente se tornando disponíveis para registros de saúde em muitos países.

Como ilustração dessa abordagem, SNPs para um haplótipo HLA-DRB1 de classe II importante (tal como descrito no Capítulo 8) foram rastreados em mais de 4.800 fenótipos em PE de mais de 4.000 pacientes; este PheWAS detectou associação não só com esclerose múltipla (como esperado a partir de estudos anteriores), mas também com cirrose hepática induzida por álcool, condições eritematosas como rosácea, várias neoplasias benignas e várias dezenas de outros fenótipos.

Embora o potencial do PheWAS esteja apenas sendo percebido, esse questionamento sem vieses de vastos conjuntos de dados pode possibilitar a descoberta de comorbidades anteriormente não avaliadas e/ou efeitos colaterais menos comuns ou interações medicamentosas em pacientes que recebem fármacos com prescrição.

## ENCONTRAR GENES RESPONSÁVEIS POR DOENÇAS POR SEQUENCIAMENTO DO GENOMA

Neste capítulo, até agora, temos nos concentrado em duas abordagens para mapear e então identificar genes envolvidos na doença, a análise de ligação e o GWAS. Agora nos voltamos para uma terceira abordagem, que envolve sequenciamento direto do genoma de indivíduos acometidos e seus pais e/ou outros indivíduos na família ou população.

O desenvolvimento de métodos amplamente melhorados de sequenciamento de DNA, que cortou o custo do sequenciamento em seis ordens de magnitude a partir do que foi gasto gerando a sequência de referência do Projeto Genoma Humano abriu novas possibilidades para descobrir genes e mutações responsáveis por doenças, especialmente no caso de distúrbios mendelianos raros. Como introduzido no Capítulo 4, estas novas tecnologias tornam possível gerar uma **sequência de genoma completo** (abreviatura em inglês, **WGS**) ou, no que pode ser um comprometimento custo-efetivo, a sequência de somente os aproximadamente 2% do genoma que contêm os exons dos genes, conhecido como sequência de exoma completo (WES).

### Filtragem da Sequência de Genoma Completo ou Dados da Sequência de Exoma Completo para Encontrar Potenciais Variantes Causais

Como exemplo do que é agora possível, considere um "trio" familiar que consiste em uma criança acometida por uma doença rara e seus pais. O WGS é realizado para todos os três, produzindo mais de 4 milhões de diferenças em comparação com a sequência referência do genoma humano (Cap. 4). Qual dessas variantes é responsável pela doença? Extrair informações úteis a partir desta enorme quantidade de dados depende de se criar um esquema de filtragem da variante com base em uma variedade de suposições razoáveis sobre que variantes são mais propensas a serem responsáveis pela doença.

Um exemplo de um esquema de filtragem que pode ser utilizado para classificar essas variantes é mostrado na Figura 10-12.

1. *Localização com relação aos genes codificantes de proteínas*. Manter variantes que estão dentro ou próximas dos exons de genes codificantes de proteínas e descartar variantes profundamente dentro de íntrons ou regiões intergênicas. É possível, é claro, que a mutação responsável possa estar em um gene de RNA não codificante ou em sequências reguladoras localizadas a alguma distância de um gene, como introduzido no Capítulo 3. Entretanto, estes são atualmente mais difíceis de avaliar e, portanto, como uma hipótese simplificadora, é razoável focar inicialmente nos genes codificantes da proteína.

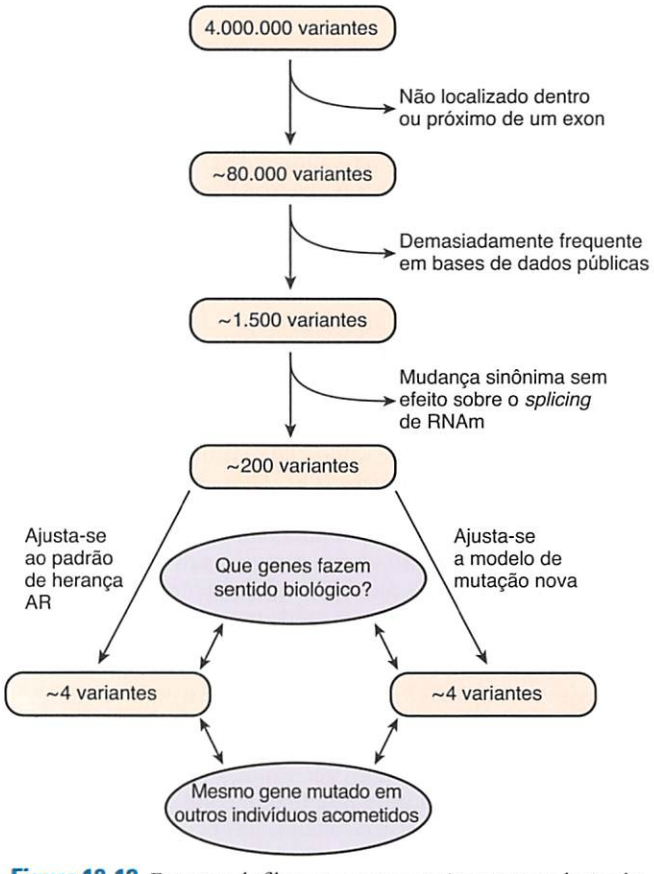



**Figura 10-12** Esquema de filtragem representativo para a redução das milhões de variantes detectadas no sequenciamento do genoma completo de uma família composta de dois progenitores não acometidos e uma criança acometida para um número pequeno que pode ser avaliado para relevância biológica e da doença. A enorme coleção inicial de variantes é selecionada em grupos cada vez menores, aplicando-se filtros que removem variantes que provavelmente não são causais com base na suposição de que as variantes de interesse são suscetíveis a estarem localizadas perto de um gene, vão perturbar sua função e são raras. Cada gene candidato restante é, em seguida, avaliado para se saber se as variantes naquele gene são herdadas de maneira que se ajuste ao padrão de herança mais provável da doença, se uma variante ocorre em um gene candidato que tem sentido biológico dado o fenótipo na criança acometida e se outros indivíduos acometidos também têm mutações naquele gene. AR, autossômico recessivo; RNAm, RNA mensageiro.

2. *Frequência populacional.* Mantenha as variantes raras da etapa 1 e descarte variantes comuns com frequências alélicas maiores que 0,05 (ou algum outro número arbitrário entre 0,01 e 0,1) porque é altamente improvável que as variantes comuns sejam responsáveis por uma doença, cuja prevalência na população seja muito menor que $q^2$, o previsto pelo equilíbrio de Hardy-Weinberg (Cap. 9).
3. *Natureza deletéria da mutação.* Mantenha variantes da etapa 2 que causam alterações *nonsense* ou não sinônimas nos códons dentro dos exons, causam mutações *frameshift* ou alteram locais de *splicing* altamente conservados, e descarte alterações sinônimas que não têm efeito previsto na função gênica.
4. *Compatibilidade com provável padrão de herança.* Se o distúrbio for considerado mais provável de ser autossômico recessivo, manter as variantes da etapa 3 que são encontradas em ambas as cópias de um gene em uma criança acometida. A criança não precisa ser homozigota para a mesma variante deletéria, mas poderia ser um heterozigoto composto para duas mutações deletérias diferentes no mesmo gene (Cap. 7). Se o modo hipotético de herança for correto, então os pais devem ambos ser heterozigotos para as variantes. Se houvesse consanguinidade nos pais, os genes candidatos e outras variantes poderiam ser filtradas, exigindo que a criança seja um verdadeiro homozigoto para a mesma mutação derivada de um único ancestral comum (Cap. 9). Se o distúrbio for grave e parecer mais provável de ser uma mutação dominante nova, porque pais não acometidos raramente ou nunca têm mais de um filho acometido, mantenha as variantes da etapa 3 que são alterações originais na criança e não estão presentes em qualquer dos progenitores.

No final, milhões de variantes podem ser filtradas até algumas poucas que ocorrem em um pequeno número de genes. Quando a filtragem reduz o número de genes e alelos para um número gerenciável, eles podem ser avaliados para outras características. Em primeiro lugar, qualquer dos genes tem uma função conhecida ou padrão de expressão do tecido que seria esperado se ele fosse o gene potencial da doença? O gene está envolvido em outros fenótipos da doença ou tem um papel nas vias com outros genes, nos quais as mutações podem causar fenótipos semelhantes ou diferentes? Finalmente, este mesmo gene é mutado em outros pacientes com a doença? Encontrar mutações em um desses genes em outros pacientes então confirmaria que este é o gene responsável no trio original.

Em alguns casos, um gene da lista na etapa 4 pode subir ao topo como um candidato porque seu envolvimento faz sentido biológico ou genético ou é conhecido como mutado em outros indivíduos acometidos. Em outros casos, no entanto, o gene responsável pode vir a ser inteiramente imprevisto em termos biológicos ou pode não ser mutado em outros indivíduos acometidos por causa da heterogeneidade de *locus* (ou seja, mutações em outros genes ainda não descobertos podem causar uma doença).

Essas avaliações de variantes requerem o uso extenso de bases de dados genômicas públicas e ferramentas de softwares. Estes incluem a sequência referência do genoma humano, bases de dados de frequências alélicas, softwares que avaliam o quão deletéria pode ser uma substituição de aminoácido para a função gênica, coleções de mutações causais de doença e bases de dados de redes funcionais e de vias biológicas. A enorme expansão dessa informação ao longo dos últimos anos tem desempenhado um papel crucial para facilitar a descoberta do gene de distúrbios mendelianos raros.

### *Exemplo*: Identificação do Gene Mutado em Disostose Acrofacial Pós-axial

A abordagem de WGS que acabou de ser delineada foi utilizada no estudo de uma família em que dois irmãos acometidos por uma malformação congênita rara conhecida como **disostose acrofacial pós-axial (DAPA)** nasceram de dois progenitores não acometidos, não aparentados. Os pacientes com esse distúrbio têm mandíbulas pequenas, ausência ou mal desenvolvimento de dígitos nos lados lunares das mãos, subdesenvolvimento da ulna, fenda labial e fendas (colobomas) das pálpebras. O distúrbio foi considerado autossômico recessivo porque os pais de uma criança acometida em algumas outras famílias são consanguíneos e existem algumas famílias, como essa aqui, com vários irmãos acometidos nascidos de pais não acometidos - ambos achados que são marcos de herança recessiva (Cap. 7). Esta pequena família isoladamente era claramente inadequada para análise de ligação. Em vez disso, todos os quatro membros da família tiveram seus genomas inteiros sequenciados e analisados.

A partir de uma lista inicial de mais de 4 milhões de variantes e supondo herança autossômica recessiva do distúrbio em ambos os filhos acometidos, um esquema de filtragem semelhante ao que foi descrito anteriormente produziu apenas quatro possíveis genes. Foi demonstrado que um destes, o *DHODH*, também era mutado em dois outros pacientes não aparentados com DAPA, confirmando assim que esse gene foi responsável pelo distúrbio nessas famílias. O *DHODH* codifica a diidroorotato desidrogenase, uma enzima mitocondrial envolvida na biossíntese de pirimidinas e não era suspeita em termos biológicos de ser o gene responsável por esta síndrome de malformação.

## Aplicações da Sequência do Genoma Completo ou Sequência do Exoma Completo em Ambientes Clínicos

Como a aplicação de WGS ou WES para distúrbios mendelianos raros foi descrita pela primeira vez em 2009, muitas centenas desses distúrbios foram estudadas e as mutações causais foram encontradas em mais de 300 genes de doenças anteriormente não reconhecidos. Embora a abordagem de sequenciamento do genoma possa negligenciar determinadas categorias de mutação que são difíceis de detectar rotineiramente por sequenciamento isolado (p. ex., deleções ou variantes do número de cópias) ou que são difíceis ou impossíveis de reconhecer com nossa compreensão atual (p. ex., mutações não codificantes ou mutações reguladoras em regiões intergênicas), muitos grupos relatam taxas de sucesso de até 25% a 40% na identificação de uma mutação causal. Essas descobertas não apenas fornecem informações úteis para o aconselhamento genético nas famílias envolvidas, mas também podem informar o manejo clínico e o potencial desenvolvimento de tratamentos eficazes.

Prevê-se que a taxa de sucesso dessa abordagem somente aumentará à medida que os custos do sequenciamento continuem caindo e nossa capacidade de interpretar as prováveis consequências funcionais das mudanças na sequência do genoma melhore.

### REFERÊNCIAS GERAIS


Altshuler D, Daly MJ, Lander ES: Genetic mapping in human disease, *Science* 322:881-888, 2008.

Manolio TA: Genomewide association studies and assessment of the risk of disease, *N Engl J Med* 363:166-176, 2010.

Risch N, Merikangas K: The future of genetic studies of complex human diseases, *Science* 273:1516-1517, 1996.

Terwilliger JD, Ott J: *Handbook of human genetic linkage*, Baltimore, 1994, Johns Hopkins University Press.


### REFERÊNCIAS PARA TÓPICOS ESPECÍFICOS


Abecasis GR, Auton A, Brooks LD, et al: An integrated map of genetic variation from 1,092 human genomes, *Nature* 491:56-65, 2012.

Bainbridge MN, Wiszniewski W, Murdock DR, et al: Whole-genome sequencing for optimized patient management, *Science Transl Med* 3, 2011, 87re3.

Bush WS, Moore JH: Genome-wide association studies, *PLoS Computational Biol* 8:e1002822, 2012.

Denny JC, Bastarache L, Ritchie MD, et al: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association data, *Nat Biotechnol* 31:1102-1110, 2013.

Fritsche LG, Chen W, Schu M, et al: Seven new loci associated with age-related macular degeneration, *Nat Genet* 17:1783-1786, 2013.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA: Human genome sequencing in health and disease, *Annu Rev Med* 63:35-61, 2012.

Hindorff LA, MacArthur J, Morales J, et al. A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed February 1, 2015.

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

Kircher M, Witten DM, Jain P, et al: A general framework for estimating the relative pathogenicity of human genetic variants, *Nat Genet* 46:310-315, 2014.

Koboldt DC, Steinberg KM, Larson DE, et al: The next-generation sequencing revolution and its impact on genomics, *Cell* 155:27-38, 2013.

Manolio TA: Bringing genome-wide association findings into clinical use, *Nat Rev Genet* 14:549-558, 2014.

Matise TC, Chen F, Chen W, et al: A second-generation combined linkage-physical map of the human genome, *Genome Res* 17:1783-1786, 2007.

Roach JC, Glusman G, Smit AF, et al: Analysis of genetic inheritance in a family quartet by whole-genome sequencing, *Science* 328:636-639, 2010.

Robinson PC, Brown MA: Genetics of ankylosing spondylitis, *Mol Immunol* 57:2-11, 2014.

Collaborative Group: SEARCH: *SLCO1B1* variants and statin-induced myopathy—a genomewide study, *N Engl J Med* 359:789-799, 2008.

Stahl EA, Wegmann D, Trynka G, et al: Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis, *Nature Genet* 44:4383-4391, 2012.

Yang Y, Muzny DM, Reid JG, et al: Clinical whole-exome sequencing for the diagnosis of mendelian disorders, *N Engl J Med* 369:1502-1511, 2013.

## PROBLEMAS

1. O *locus* da doença de Huntington (DH) foi considerado rigidamente ligado a um polimorfismo de DNA no cromossomo 4. No mesmo estudo, no entanto, foi descartada ligação entre DH e o *locus* para o polimorfismo de grupo sanguíneo MNS, que também está mapeado no cromossomo 4. Qual é a explicação?

2. LOD *scores* (Z) entre um polimorfismo no *locus* de α-globina no braço curto do cromossomo 16 e uma doença autossômica dominante foram analisados em uma série de famílias britânicas e holandesas, com os seguintes dados:

| θ | 0,00 | 0,01 | 0,10 | 0,20 | 0,30 | 0,40 |
|---|---|---|---|---|---|---|
| Z | -∞ | 23,4 | 24,6 | 19,5 | 12,85 | 5,5 |

$Z_{máx} = 25,85$ em $\theta_{max} = 0,05$

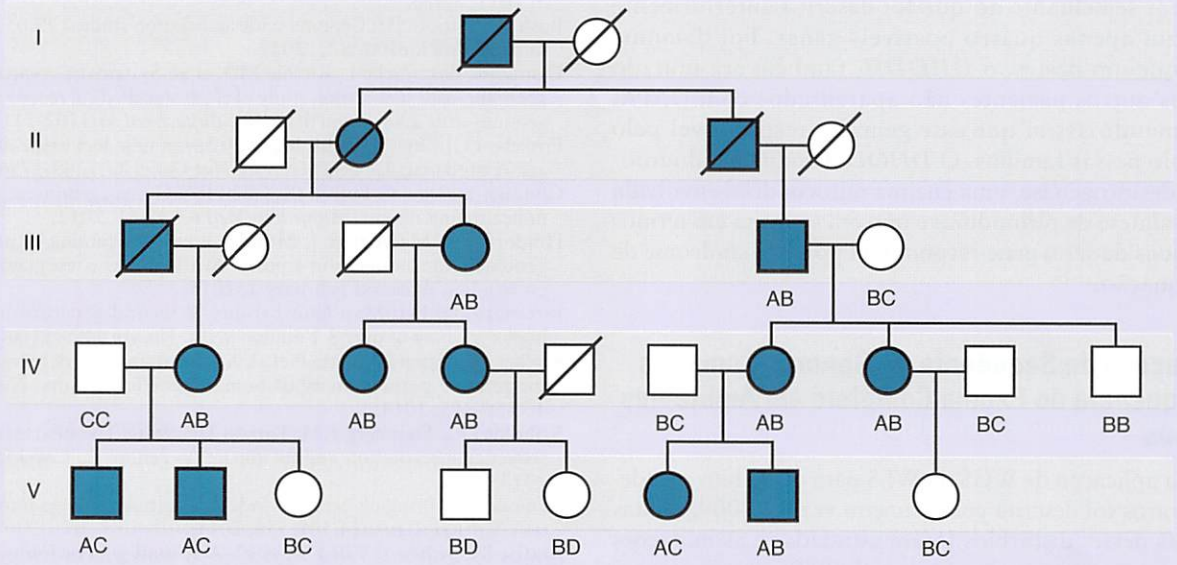
Como você interpretaria estes dados? Por que o valor dado de Z como -∞ em θ = 0?

Em um estudo posterior, uma grande família da Sicília com o que parece ser a mesma doença foi também investigada para ligação com a α-globina, com os seguintes resultados:

| θ | 0,00 | 0,10 | 0,20 | 0,30 | 0,40 |
|---|---|---|---|---|---|
| LOD *scores* (Z) | -∞ | -8,34 | -3,34 | -1,05 | -0,02 |

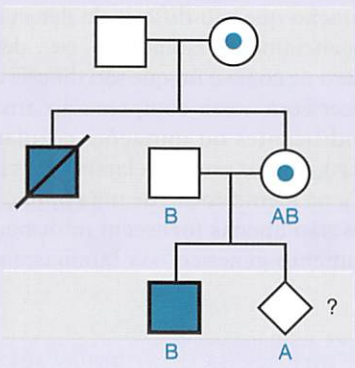
Como você interpretaria os dados neste segundo estudo?

3. Este heredograma foi obtido em um estudo projetado para determinar se uma mutação em um gene para γ-cristalina, uma das principais proteínas do cristalino ocular, pode ser responsável por uma forma de catarata autossômica dominante. Os símbolos preenchidos no heredograma indicam membros da família com catarata. As letras indicam três alelos no *locus* de γ-cristalina polimórfica no cromossomo 2. Se você examinar cada pessoa acometida que passou a catarata para seus filhos, quantos destes representam uma meiose que é informativa para ligação entre a catarata e a γ-cristalina? Em que indivíduos é conhecida a fase entre a mutação da catarata e os alelos de γ-cristalina? Há algumas meioses em que um *crossover* deve ter ocorrido para explicar os dados? O que você concluiria sobre a ligação entre a catarata e a γ-cristalina deste estudo? Que estudos adicionais podem ser realizados para confirmar ou rejeitar a hipótese?




Heredograma para questão 3

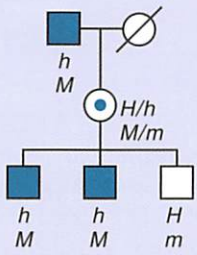
4. O seguinte heredograma mostra um exemplo de diagnóstico molecular na síndrome de Wiskott-Aldrich, uma imunodeficiência ligada ao X, por utilização de um polimorfismo de DNA ligado com uma distância de mapa de cerca de 5 cM entre o *locus* polimórfico e o gene da síndrome de Wiskott-Aldrich.
   a. Qual é a fase provável na mãe portadora? Como você determinou isso? Que diagnóstico você faria com relação ao diagnóstico pré-natal se fosse um feto do sexo masculino?
   b. O avô materno torna-se agora disponível para teste de DNA e apresenta o alelo B no *locus* ligado. Como esse achado afeta sua determinação de fase na mãe? Que diagnóstico você faria agora no que diz respeito ao diagnóstico pré-natal atual?




Heredograma para questão 4

5. Revise o heredograma na Figura 10-10B. Se a avó acometida, I-2, fosse um heterozigoto *A/a*, seria possível determinar a fase no progenitor acometido, o indivíduo II-2?

6. No hereodgrama adiante, que mostra uma família com hemofilia A ligada ao X, você consegue determinar a fase do gene do fator VIII mutante (*h*) e o alelo normal (*H*) no que diz respeito a alelos polimórficos M e *m* na mãe de dois meninos acometidos?




Heredograma de hemofilia ligada ao X. O avô acometido na primeira geração tem a doença (alelo h mutante) e alelo M em um *locus* polimórfico no cromossomo X.

7. Calcule D' para os três cenários listados na Figura 10-7.

8. Cálculos de risco relativo são usados para estudos de coorte e estudos não caso-controle. Para demonstrar por que este é o caso, imagine um estudo caso-controle para o efeito de uma variante genética na suscetibilidade à doença. O pesquisador verificou o maior número de indivíduos acometidos possível (a + c) e, em seguida, escolhe arbitrariamente um conjunto de controles (b + d). Eles são genotipados quanto a saber se uma variante está presente: a/(a + c) dos acometidos têm a variante, enquanto b/(b + d) dos controles têm a variante.

| | Doença Presente | Doença Ausente |
|---|---|---|
| Variante presente | a | b |
| Variante ausente | c | d |
| | a + c | b + d |

Calcule a *odds ratio* e o risco relativo para a associação entre a variante estar presente e a doença estar presente.

Agora, imagine que o pesquisador decidiu arbitrariamente usar três vezes mais indivíduos não acometidos, 3 x (b +d) como controles. O pesquisador tem todo o direito de fazê-lo porque é um estudo caso-controle e os números de acometidos e não acometidos não são determinados pela prevalência da doença na população que está sendo estudada, como seriam em um estudo de coorte. Suponha que a distribuição da variante permanece a mesma neste grupo controle, como no grupo controle menor que é, 3b/[3 x (b + d)] = b/(b + d) portador do alelo.

| | Doença Presente | Doença Ausente |
|---|---|---|
| Variante presente | a | 3b |
| Variante ausente | c | 3d |
| | a + c | 3 x (b + d) |

Recalcule a OR e o RR com este novo grupo controle.

Faça o mesmo quando um grupo controle arbitrário for uma n-tupla do grupo controle original; ou seja, o tamanho do grupo controle é n x (b + d).

Qual destas medidas, OR ou RR, não muda quando grupos controle diferentes, de tamanho arbitrário são usados?